MAKALAH DATA MINING
PERBANDINGAN AKURASI REGRESI LOGISTIK BINER DAN ANALISIS DISKRIMINAN PADA DATA BANKLOAN

Dosen Pengampu:
Ardiana Alifatus Sa'adah, S.Si., M.Si.

Disusun Oleh:

Zahra Ulaya Sifa	24050121130097
Hilda Ayu Meylia	24050121140119
Sekar Fajrina	24050121140121

DATA MINING – D

DEPARTEMEN STATISTIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO
SEMARANG

2024

KATA PENGANTAR

Puji syukur kepada Allah SWT atas limpahan rahmat dan nikmat-Nya, penulis diberi kemudahan dalam menyelesaikan makalah tepat waktu. Tanpa rahmat dan pertolongan-Nya, penulis tidak akan mampu menyelesaikan makalah ini dengan baik. Tidak lupa shalawat serta salam tercurahkan kepada Nabi Muhammad SAW yang syafaatnya kita nantikan kelak. Adapun tujuan dari penulisan makalah ini adalah guna memenuhi tugas mata kuliah Data Mining.

Penulis mengucapkan terima kasih, kepada Ibu Ardiana Alifatus Sa'adah, S.Si., M.Si. selaku dosen mata kuliah Data Mining yang telah memberikan bimbingan dan kepercayaan dalam menyelesaikan makalah ini. Penulis juga berterima kasih kepada pihak-pihak yang telah membantu penulis dalam penyusunan makalah ini. Penulis berharap makalah ini dapat menjadi bahan referensi yang bermanfaat.

Penulis menyadari makalah ini masih terdapat kekurangan dan jauh dari kata sempurna. Oleh karena itu, kritik dan saran pembaca senantiasa dapat diterima penulis, agar makalah ini dapat menjadi lebih baik.

Demikian yang dapat penulis sampaikan. Akhir kata, semoga makalah ini dapat bermanfaat bagi semua pihak yang membutuhkan.

Semarang, 17 Juni 2024

Penulis

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam dunia keuangan, pengambilan keputusan yang tepat sangat penting bagi keberhasilan perusahaan, terutama dalam konteks pemberian pinjaman. Pemberian pinjaman yang tidak terkelola dengan baik dapat menyebabkan masalah likuiditas dan mengganggu stabilitas keuangan perusahaan. Salah satu risiko utama dalam pemberian pinjaman adalah kredit macet, di mana peminjam gagal membayar pinjaman sesuai dengan kesepakatan awal. Dalam upaya mengelola risiko ini, bank melakukan analisis kelayakan terhadap calon peminjam yang melibatkan evaluasi mendalam terhadap faktor-faktor seperti riwayat kredit, pendapatan, pekerjaan, usia, pendidikan, jumlah tanggungan, aset yang dimiliki, dan faktor-faktor lain yang dapat mempengaruhi kemampuan peminjam untuk membayar kembali pinjaman. Selain itu, faktor eksternal seperti kondisi ekonomi juga dipertimbangkan dalam proses ini.

Dalam era yang semakin mengandalkan data, pendekatan *data-driven* sangat diperlukan dalam proses pengambilan keputusan di sektor perbankan. Thomas, Edelman, dan Crook (2002) menegaskan bahwa penggunaan teknik statistik dalam pemodelan risiko kredit telah menjadi standar industri. Dua metode yang umum digunakan adalah regresi logistik biner dan analisis diskriminan.

Regresi logistik biner memperkirakan kemungkinan terjadinya suatu peristiwa, seperti memilih atau tidak memilih, berdasarkan kumpulan data variabel independen tertentu. Menurut Hosmer dan Lemeshow (2000), regresi logistik biner adalah salah satu metode statistik yang biasanya digunakan untuk menjelaskan hubungan antara variabel respon yang bersifat biner (dikotomus) dengan satu atau lebih variabel prediktor yang bersifat metrik atau non-metrik. Variabel respon terdiri dari dua kategori, yaitu “sukses” dan “gagal” yang dinotasikan dengan 0 (sukses) dan 1 (gagal). Regresi logistik biner juga memungkinkan interpretasi yang mudah dari koefisien model, yang dapat diartikan sebagai pengaruh masing-masing variabel prediktor terhadap probabilitas terjadinya peristiwa.

Sementara itu, analisis diskriminan adalah teknik yang digunakan untuk memisahkan dua atau lebih kelompok berdasarkan kombinasi linear atau kuadratik dari variabel-variabel independen. Menurut Johnson & Wichern (2007), analisis diskriminan

adalah teknik multivariat yang digunakan untuk memisahkan objek ke dalam kelompok yang berbeda serta untuk mengklasifikasikan objek baru ke dalam salah satu kelompok yang telah ditentukan sebelumnya. Analisis diskriminan dapat memberikan insight yang mendalam tentang struktur data dan faktor-faktor yang paling mempengaruhi klasifikasi. Dalam konteks pinjaman pribadi, analisis diskriminan dapat digunakan untuk mengklasifikasikan calon peminjam ke dalam kelompok yang kemungkinan besar akan mengembalikan pinjaman dan kelompok yang kemungkinan tidak akan mengembalikan pinjaman.

Perbandingan akurasi antara regresi logistik biner dan analisis diskriminan penting dalam menentukan metode yang paling efektif untuk memprediksi kelayakan penerima pinjaman. Keakuratan prediksi ini tidak hanya membantu bank mengurangi risiko kredit macet, tetapi juga meningkatkan efisiensi operasional serta profitabilitas mereka. Lebih lanjut, kepercayaan peminjam terhadap lembaga keuangan juga dipengaruhi oleh proses penilaian yang adil dan transparan terhadap peminjam.

Berdasarkan uraian di atas, penulis tertarik untuk meneliti penerapan model regresi logistik biner dan analisis diskriminan di bidang perbankan, yaitu membandingkan kinerja kedua metode tersebut dalam memprediksi kelayakan peminjam. Sehingga terbentuk judul penelitian "Perbandingan Akurasi Regresi Logistik Biner dan Analisis Diskriminan pada Data *Bankloan*".

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang di atas, dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana akurasi regresi logistik biner dalam memprediksi kelayakan pinjaman pada data *bankloan*?
2. Bagaimana akurasi analisis diskriminan dalam memprediksi kelayakan pinjaman pada data *bankloan*?
3. Metode manakah yang lebih akurat dalam konteks data *bankloan* berdasarkan prediksi kelayakan pinjaman yang dihasilkan antara regresi logistik biner dan analisis diskriminan?

1.3 Tujuan

Berdasarkan rumusan masalah di atas, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Mengevaluasi akurasi regresi logistik biner dalam prediksi kelayakan pinjaman pada data *bankloan*.
2. Mengevaluasi akurasi analisis diskriminan dalam prediksi kelayakan pinjaman pada data *bankloan*.
3. Membandingkan akurasi regresi logistik biner dan analisis diskriminan, serta menentukan metode yang lebih efektif dalam konteks data *bankloan*.

BAB II
TINJAUAN PUSTAKA

2.1 Bank *Loan*

Bank *loan* atau pinjaman bank merupakan salah satu instrumen utama dalam aktivitas perbankan modern yang memfasilitasi akses modal bagi individu, bisnis, dan entitas lainnya. Secara umum, bank *loan* merujuk kepada pinjaman yang diberikan oleh lembaga keuangan kepada peminjam dengan syarat-syarat tertentu, seperti suku bunga, jangka waktu, dan jaminan. Pinjaman ini terbagi dalam beberapa jenis utama, termasuk pinjaman konsumen yang meliputi hipotek dan pinjaman kendaraan, pinjaman bisnis untuk modal kerja atau investasi, serta pinjaman komersial untuk proyek besar seperti infrastruktur. Proses pemberian pinjaman dimulai dengan evaluasi kredit yang ketat, di mana bank melakukan analisis mendalam terhadap kelayakan peminjam. Evaluasi ini meliputi penilaian terhadap kredit skor peminjam, riwayat kredit, serta kemampuan dan kestabilan finansial untuk memenuhi kewajiban pembayaran pinjaman. Selain dari evaluasi kredit konvensional, bank *loan* juga mempertimbangkan faktor-faktor tambahan seperti umur, pengalaman kerja, pendapatan, dan variabel demografis seperti kode ZIP dan jumlah anggota keluarga. Faktor-faktor ini memberikan gambaran yang lebih komprehensif tentang profil risiko peminjam serta kemampuannya untuk mengelola pinjaman dengan baik. Pengalaman kerja yang lebih lama atau tingkat pendidikan yang lebih tinggi sering kali mengindikasikan stabilitas finansial yang lebih kokoh, sementara rata-rata penggunaan kartu kredit (CCAvg) memberikan petunjuk tentang kebiasaan pengeluaran dan potensi kemampuan peminjam untuk melakukan pembayaran tepat waktu. Kondisi hipotek yang dimiliki juga dapat memberikan gambaran tentang beban utang peminjam dan ketersediaan sumber daya finansial mereka. Selain itu, kepemilikan rekening sekuritas (Securities Account) atau akun deposito (CD Account) mencerminkan adanya aset lain yang bisa digunakan sebagai jaminan atau menunjukkan kemampuan peminjam dalam mengelola investasi secara bijak. Faktor-faktor seperti transaksi *online* (Online) dan kepemilikan kartu kredit (CreditCard) juga penting dalam memahami perilaku keuangan peminjam serta kecenderungan penggunaan produk perbankan modern. Keseluruhan informasi ini, ketika dipertimbangkan bersama, memberikan penilaian yang lebih holistik terhadap risiko kredit dan mendukung keputusan pemberian pinjaman yang lebih tepat bagi pihak bank.

Bank *loan* tidak hanya merupakan sarana untuk memperoleh modal, tetapi juga membawa sejumlah risiko yang perlu dikelola dengan cermat. Risiko kredit menjadi salah satu yang paling signifikan, di mana ketidakmampuan peminjam untuk membayar pinjaman dapat berakibat pada kerugian finansial bagi bank. Faktor risiko ini diperparah oleh fluktuasi kondisi ekonomi dan suku bunga, yang dapat mempengaruhi kemampuan peminjam untuk memenuhi kewajiban finansial mereka. Selain risiko kredit, bank *loan* juga rentan terhadap risiko suku bunga, khususnya untuk pinjaman dengan suku bunga variabel atau jangka waktu panjang. Meskipun demikian, bank *loan* juga memberikan manfaat yang signifikan bagi peminjam, termasuk akses modal untuk investasi atau kebutuhan mendesak, serta pembangunan riwayat kredit yang baik yang dapat mendukung kegiatan finansial masa depan.

Perkembangan dalam teknologi dan regulasi telah mengubah lanskap industri perbankan, terutama dalam praktik pemberian pinjaman. Adopsi teknologi *fintech*, seperti platform pinjaman *online* dan penggunaan *big* data untuk analisis risiko kredit, telah memungkinkan bank untuk meningkatkan efisiensi dalam proses pemberian pinjaman dan memberikan pengalaman yang lebih personal kepada nasabah. Teknologi *blockchain*, misalnya, memberikan tingkat transparansi dan keamanan yang lebih tinggi dalam transaksi pinjaman, yang dapat meningkatkan kepercayaan antara peminjam dan pemberi pinjaman. Selain itu, perubahan dalam regulasi perbankan juga memainkan peran penting dalam menentukan strategi pemberian pinjaman bank, termasuk persyaratan modal minimum dan kebijakan suku bunga. Pemahaman mendalam tentang faktor-faktor ini tidak hanya penting bagi bank dalam mengelola risiko dan mengambil keputusan bisnis yang tepat, tetapi juga untuk memahami dampaknya terhadap stabilitas ekonomi makro dan kesejahteraan finansial masyarakat secara keseluruhan.

2.2 Preprocessing Data

Preprocessing data adalah tahapan penting dalam analisis data dan *machine learning* yang bertujuan untuk mempersiapkan data mentah agar dapat digunakan secara efektif. Berikut tahapan *preprocessing* yang akan digunakan dalam penelitian kali ini:

1. **Pemindahan Kolom Variabel Dependent:** Langkah awal adalah memastikan bahwa kolom variabel *dependent* (y) dipindahkan ke posisi terakhir dalam *dataset*. Hal ini dilakukan agar mempermudah analisis serta memastikan bahwa data yang digunakan memiliki struktur yang sesuai.

2. **Pengecekan dan Penghapusan Missing Values**: Dilakukan pengecekan terhadap keberadaan nilai yang hilang *(missing values)* dalam *dataset*. Jika ditemukan *missing values,* baris data yang terdapat nilai yang hilang akan dihapus. Langkah ini penting untuk memastikan konsistensi dan validitas *dataset* yang akan digunakan dalam analisis.
3. **Penghapusan Outlier**: *Outlier* atau pencilan merupakan nilai yang signifikan dari pola umum data. Penghapusan outlier dilakukan untuk mencegah nilai-nilai ekstrem ini mempengaruhi analisis secara tidak proporsional. Metode yang umum digunakan adalah dengan menghitung *Interquartile Range* (IQR) untuk setiap kolom dan menghapus data yang dianggap sebagai *outlier.*
4. **Evaluasi Standar Deviasi**: Standar deviasi dari setiap kolom dievaluasi untuk mengidentifikasi kolom-kolom yang memiliki variabilitas rendah atau bahkan nol. Kolom-kolom seperti ini cenderung tidak memberikan informasi yang signifikan dalam analisis dan dapat dihapus dari *dataset.*
5. **Evaluasi Korelasi**: Matriks korelasi antar variabel dievaluasi untuk mengidentifikasi dan mengatasi masalah *multicollinearity.* Variabel yang memiliki korelasi tinggi dapat menyebabkan masalah dalam model analisis, karena informasi yang sebenarnya dapat terduplikasi atau terlalu tergantung pada satu sama lain.
6. **Evaluasi IQR (Interquartile Range)**: IQR digunakan untuk mengevaluasi sebaran nilai dalam setiap kolom. Kolom-kolom dengan IQR yang sangat kecil, atau bahkan nol, menunjukkan bahwa data dalam kolom tersebut tidak memiliki variasi yang cukup signifikan untuk memberikan kontribusi yang bermakna dalam analisis.

2.3 Regresi Logistik Biner

Model regresi logistik biner digunakan untuk menganalisis hubungan antara satu variabel respon (y) dan beberapa variabel bebas (x). Adapun variabel responnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan sebuah karakteristik dan bernilai 0 untuk menyatakan ketidakberadaan sebuah karakteristik. Jika diketahui Y variabel respon bernilai 1 dan 0, maka:

$$\mathrm{P}(\mathrm{Y}=1|\mathrm{X}=x_i)=\pi(x_i) \text{ dan } \mathrm{P}(\mathrm{Y}=0|\mathrm{X}=x_i)=1-\pi(x_i)$$

Dengan:

- $\mathrm{i}=1,2,..,\mathrm{p}$
- $\mathrm{P}(\mathrm{Y}=1|\mathrm{X}=x_i)$: probabilitas bahwa variabel respon (Y) bernilai 1 diberikan variabel independen X

Sehingga model regresi logistik biner: $\pi(x_i) = \frac{exp\{g(x_i)\}}{1 + exp\{g(x_i)\}}$

Dengan $g(x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$

Sedangkan logit dari $\pi(x_i)$ adalah:

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

dengan:

- β_0 : *intercept*
- $\beta_1, \beta_2, \ldots, \beta_p$: koefisien regresi logistik biner untuk variabel bebas $x_1, x_2, \ldots, x_p$
- Exp : adalah basis logaritma natural (sekitar 2.71828)

Estimasi dari parameter diperoleh dengan menggunakan metode maksimum Likelihood yang selanjutnya diselesaikan dengan metode iterasi Newton Raphson. Terdapat beberapa uji yang dilakukan untuk menentukan parameter yang signifikan untuk model, yaitu sebagai berikut:

1. Uji Rasio Likelihood

Uji yang digunakan untuk membandingkan model yang mengandung variabel bebas dan model yang tidak mengandung variabel bebas. Uji ini digunakan apakah model signifikan atau tidak.

- Hipotesis

  $H_0: \beta_1 = \beta_{\ldots} = \beta_p = 0$ (Secara bersama-sama variabel bebas tidak memengaruhi model)

  H_1: Salah satu dari $\beta_k \neq 0$ dengan k=1,2,.. (Secara bersama-sama variabel bebas memengaruhi model)

- Taraf Signifikansi: α
- Statistik Uji

  $G = -2 \ln\left(\frac{Likelihood\ tanpa\ variabel\ bebas}{Likelihood\ dengan\ variabel\ bebas}\right)$

- Kriteria Uji

  Tolak H_0 jika nilai $G > X^2_{(\alpha,p)}$ atau p-value < α

2. Uji Goodnes of Fit

Uji Goodness of Fit dalam regresi logistik biner bertujuan untuk menilai sejauh mana model logistik sesuai dengan data yang diamati. Dalam konteks regresi logistik biner, tujuan utamanya adalah untuk memastikan bahwa model yang dibangun dengan baik dapat memprediksi probabilitas kejadian secara akurat

berdasarkan variabel independen yang digunakan. Tes Hosmer-Lemeshow adalah salah satu uji GOF yang paling umum digunakan untuk regresi logistik biner. Tes ini menguji apakah ada perbedaan yang signifikan antara frekuensi kejadian yang diamati dan yang diprediksi oleh model.

- Hipotesis

  H_0: Model sesuai (observasi dan prediksi tidak berbeda)

  H_1: Model tidak sesuai (observasi dan prediksi berbeda)
- Taraf Signifikansi: α
- Statistik Uji

  Hosmer-Lemeshow Test:

  $C = \sum_{k=1}^{g} \frac{(O_k - n_k^{`}\overline{\pi_k})^2}{n_k^{`}\overline{\pi_k}(1-\overline{\pi_k})}$

  dengan:

  O_k: Jumlah nilai variabel respon pada grup ke-k

  $\overline{\pi_k}$: Rata-rata taksiran peluang pada grup ke-k

  g : Banyak grup

  $n_k^{`}$: Banyak observasi pada grup ke-k
- Kriteria Uji

  Tolak H_0 jika nilai C > $X^2_{(\alpha,g-2)}$ atau p-value < α

3. Uji Wald

Uji Wald digunakan untuk menilai signifikansi koefisien regresi logistik biner. Uji ini membantu menentukan apakah variabel bebas *(independent)* secara individual memiliki pengaruh yang signifikan terhadap variabel respon *(dependent)*. Dalam konteks regresi logistik biner, tujuan utama dari uji Wald adalah untuk menguji hipotesis nol bahwa koefisien suatu variabel adalah nol (tidak berpengaruh).

- Hipotesis

  H_0: $\beta_j = 0$, j=1,2,..,j (Variabel bebas tidak signifikan terhadap model)

  H_1: $\beta_j \neq 0$, j=1,2,..,j (Variabel bebas signifikan terhadap model)
- Taraf Signifikansi: α
- Statistik Uji

  $Wj = \left(\frac{\widehat{\beta_j}}{Se\widehat{\beta_j}}\right)^2$
- Kriteria Uji

  Tolak H0 jika w_j > $\chi^2_{(\alpha,1)}$ atau p-value < α

2.4 Analisis Diskriminan

Analisis Diskriminan adalah salah satu teknik analisa statistika dependensi yang memiliki kegunaan untuk mengklasifikasikan objek ke beberapa kelompok. Pengelompokan dengan analisis diskriminan ini terjadi karena ada pengaruh satu atau lebih variabel lain yang merupakan variabel independen. Kombinasi linier dari variabel-variabel ini akan membentuk suatu fungsi diskriminan (Tatham et. al.,1998). Analisis diskriminan linier (LDA), analisis diskriminan normal (NDA), atau analisis fungsi diskriminan adalah generalisasi dari diskriminan linier fisher, merupakan suatu metode yang digunakan dalam statistic untuk menemukan kombinasi fitur linier yang mencirikan atau memisahkan dua kelas atau lebih dari satu objek atau peristiwa.

Analisis diskriminan memiliki variabel independen kontinu dan variabel dependen kategoris (yaitu label kelas) dan digunakan ketika kelompok diketahui secara apriori, dengan setiap kasus harus memiliki skor pada satu atau lebih ukuran *predictor* kuantitatif, dan skor pada ukuran kelompok. Secara sederhana, analisis diskriminan adalah klasifikasi atau tindakan mendistribusikan sesuatu ke dalam kelompok, kelas, atau kategori yang sejenis.

Model analisis diskriminan adalah sebuah persamaan yang menunjukkan suatu kombinasi linier dari berbagai variabel independen. Dengan model sebagai berikut:

$$D = b_0 + b_1X_1 + b_2X_2 + \cdots b_kX_k$$

dimana:

D = Skor diskriminan

b_0 = Konstanta

b_i = Koefisien diskriminan atau bobot

X_k = Prediktor atau variabel *independent*

Analisis diskriminan terdiri dari lima tahapan analisis, yaitu (1) merumuskan masalah, (2) mengestimasi koefisien fungsi diskriminan, (3) memastikan signifikansi determinan, (4) menginterpretasi hasil, dan (5) menguji signifikansi analisis diskriminan (Malhotra 2006).

2.4.1 Asumsi Analisis Diskriminan

Asumsi yang harus dipenuhi dalam analisis diskriminan sama dengan asumsi MANOVA. Analisis ini cukup sensitive terhadap outlier dan ukuran kelompok terkecil harus lebih besar dari jumlah variabel *predictor*, asumsi analisis

diskriminan diantaranya adalah normal multivariat, homogenitas varians/kovarians (homokedastisitas), dan non multikolinieritas.

- Normal Multivariat

  Pengecekan asumsi normal multivariat dapat dilihat dari Q-Q plot antara square distance $\left(d_j{}^2\right)$ dengan nilai quantil dari distribusi Chi-Square $\left(\frac{j-0.5}{n}\right)$. Jika hasil plot menggambarkan garis lurus maka data tersebut dapat dinyatakan sebagai normal multivariat atau dilakukan dengan pengujian formal yang dirumuskan sebagai berikut:

  - Hipotesis

    H_0: Data berdistribusi normal multivariat

    H_1: Data tudak berdistribusi normal multivariat

  - Taraf Signifikansi

    $\alpha = 5\%$

  - Statistik uji:

$$r_q = \frac{\sum_{j-1}^{n}(x_j - \bar{x})(q_j - \bar{q})}{\sqrt{\sum_{j-1}^{n}(x_j - \bar{x})^2}\sqrt{\sum_{j-1}^{n}(q_j - \bar{q})^2}}$$

  - Kriteria Uji

    Tolak H_0 jika $r_q < r_{n,a}$ atau nilai p-value < α

  Namun, analisis diksriminan relative kuat terhadap pelangggaran terhadap asumsi ini, dan juga dikatakan bahwa analisis diskriminan mungkin masih dapat diandalkan ketika menggunakan variabel dikotomi, dimana asumsi normal multivariat sering dilanggar.

- Homogenitas Varians atau Kovarians (Homoskedastisitas)

  Uji homogenitas adalah pengujian mengenai sama tidaknya variansi-variansi dua buah distribusi atau lebih, terlihat dari hasil pengujian statistic Box's M. Namun, analisis diskriminan dianjurkan digunakan ketika kovariansnya sama, sedangkan analisis diskriminan kuadrat digunakan ketika kovariansnya tidak sama. Pemeriksaan kesamaan matriks varians kovarians antara dua populasi atau lebih dilakukan dengan Box's M test yang dirumuskan sebagai berikut.

  - Hipotesis

$H_0 : \sum_1 = \sum_2 = ... \sum_k = \sum$ (matriks kovarians bersifat multivariat homoskedastisitas)

H_1 : Minimal ada satu $\sum_i \neq \sum_j$ (matriks kovarians tidak bersifat multivariat homoskedastisitas

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik uji

$$S_x^2 = \sqrt{\frac{n.\sum X 2-(\sum x)2}{n(n-1)}} \qquad Sr^2 = \sqrt{\frac{n.\sum Y 2 - (\sum Y)2}{n(n-1)}}$$

$$F = \frac{S_{besar}}{S_{kecil}}$$

- Kriteria Uji

  Tolak H_0 jika nilai p-value < α

2.4.2 Uji Wilk Lambda

Uji Wilk Lamnda merupakan pengujian yang dilakukan untuk menguji variabel manakah yang memberikan kontribusi signifikan dalam fungsi diskriminat. Semakin dekat nilai lambda wilk dengan 0, maka semakin besar kontribusi terhadap fungsi diskriminan. Adapun nilai statistic chi-square yang dihasilkan untuk menguji signifikansi lambda wilk, jika nilai p-value kurang dari 0.05 maka fungsi terkait menjelaskan keanggotaan kelompok dengan baik.

2.5 Perbandingan Regresi Logistik Biner dan Analisis Diskriminan

Regresi logistik biner dan analisis diskriminan merupakan dua teknik statistik yang digunakan untuk klasifikasi dan prediksi dalam berbagai bidang, seperti biomedis, pemasaran, dan ilmu sosial. Meskipun keduanya digunakan untuk memprediksi hasil kategoris, pendekatan dan asumsi yang mendasari keduanya berbeda.

Regresi logistik biner menggunakan model berbasis probabilitas untuk memprediksi kelas dari variabel dependen biner. Model ini mengasumsikan bahwa log odds dari variabel dependen adalah fungsi linear dari variabel independen. Salah satu keuntungan utama regresi logistik biner adalah fleksibilitas, model ini tidak memerlukan asumsi distribusi normal dari variabel independen dan dapat dengan mudah diperluas untuk menangani kasus-kasus dengan lebih dari dua kategori melalui regresi logistik biner multinomial.

Di sisi lain, analisis diskriminan, khususnya analisis diskriminan linier (LDA), mengasumsikan bahwa variabel independen memiliki distribusi normal multivariat dengan kovarians yang sama di setiap kelompok. LDA membangun fungsi diskriminan yang merupakan kombinasi linear dari variabel independen yang memaksimalkan rasio antara varian antar-kelompok terhadap varian dalam-kelompok, sehingga memisahkan kelompok-kelompok dengan sebaik mungkin.

Regresi logistik biner tidak memberikan hasil yang baik jika terdapat multikolinearitas tinggi di antara variabel independen. Sementara itu, LDA sangat sensitif terhadap pelanggaran asumsi normalitas dan homogenitas kovarians, yang dapat mempengaruhi akurasi klasifikasi. Secara keseluruhan, pemilihan antara regresi logistik biner dan analisis diskriminan bergantung pada sifat data dan asumsi yang dapat dipenuhi. Regresi logistik biner lebih fleksibel dalam hal asumsi distribusi, sedangkan LDA bisa lebih efisien dalam situasi dengan data normal multivariat dan homogenitas kovarians. Kedua metode memiliki tempatnya masing-masing dalam analisis statistik, dan pemahaman yang mendalam tentang asumsi dan karakteristik data akan membantu menentukan metode yang paling tepat untuk digunakan.

2.5.1 Asumsi Regresi Logistik Biner dan Analisis Diskriminan

Regresi logistik biner dan analisis diskriminan memiliki asumsi-asumsi yang berbeda yang perlu dipenuhi agar hasil analisisnya terpercaya. Berikut ini adalah penjelasan mengenai asumsi-asumsi dari masing-masing metode:

A. Regresi logistik biner

- Independensi kesalahan: Observasi harus independen satu sama lain. Ini berarti kesalahan atau residu dari satu observasi tidak boleh berkorelasi dengan kesalahan dari observasi lain.
- Tidak ada multikolinearitas: Multikolinearitas tinggi antara variabel independen harus dihindari. Artinya, variabel-variabel independen tidak boleh memiliki korelasi yang sangat tinggi satu sama lain.
- Biner atau multikategori variabel dependen: variabel dependen dalam sebuah analisis harus berupa variabel kategoris. Variabel tersebut bisa dalam bentuk biner, yang hanya memiliki dua kategori (misalnya, ya atau tidak, sukses atau gagal), atau dalam bentuk multinomial, yang memiliki lebih dari dua kategori (misalnya, merah, biru, hijau).

B. Analisis Diskriminan

- Normalitas multivariat: Variabel-variabel independen harus mengikuti distribusi normal multivariat dalam setiap kelompok dari variabel dependen. Ini berarti bahwa dalam setiap kategori dari variabel dependen, kombinasi dari variabel-variabel independen harus membentuk distribusi yang normal.
- Homogenitas kovarians: Matriks kovarians dari variabel-variabel independen harus sama di setiap kelompok dari variabel dependen. Ini berarti bahwa varians dan kovarians antar variabel-variabel prediktor harus tetap konsisten di seluruh kelompok. Dengan kata lain, pola variabilitas dan hubungan antar variabel independen tidak boleh berubah tergantung pada kategori dari variabel dependen.

2.5.2 Kelebihan dan Kelemahan Regresi Logistik Biner

Kelebihan regresi logistik biner diantaranya adalah sebagai berikut:

- Regresi logistik biner bekerja dengan baik ketika kumpulan data dapat dipisahkan secara linier**.**
- Regresi logistik biner tidak terlalu rentan terhadap over-fitting namun dapat mengalami overfit pada kumpulan data berdimensi tinggi. Anda harus mempertimbangkan teknik regularisasi (L1 dan L2) untuk menghindari penyesuaian yang berlebihan dalam skenario ini.
- Regresi logistik biner tidak hanya memberikan ukuran seberapa relevan suatu prediktor (ukuran koefisien), namun juga arah asosiasinya (positif atau negatif).
- Regresi logistik biner lebih mudah diimplementasikan, diinterpretasikan, dan sangat efisien untuk dilatih.

Kelemahan regresi logistik biner diantaranya adalah sebagai berikut:

- Batasan utama regresi logistik biner adalah asumsi linearitas antara variabel terikat dan variabel bebas. Di dunia nyata, data jarang dapat dipisahkan secara linier sehingga seringkali data menjadi berantakan.
- Jika jumlah observasi lebih sedikit dari jumlah fitur, regresi logistik biner tidak boleh digunakan, karena dapat menyebabkan overfit.
- Regresi logistik biner hanya dapat digunakan untuk memprediksi fungsi diskrit. Oleh karena itu, variabel terikat regresi logistik biner dibatasi pada

kumpulan angka diskrit. Pembatasan ini sendiri bermasalah karena menghambat prediksi data berkelanjutan.

2.5.3 Kelebihan dan Kelemahan Analisis Diskriminan

Kelebihan analisis diskriminan diantaranya adalah sebagai berikut:

- Efektif dalam mengelompokkan data ke dalam kategori yang sudah ditentukan dengan baik.
- Hasil analisis dapat diinterpretasikan dengan relative mudah untuk dipahami.
- Dapat mengidentifikasi variabel penting yang paling signifikan dalam memisahkan kelompok.
- Efektif digunakan saat terdapat banyak variabel independen.
- Mampu memberikan prediksi yang baik.

Kekurangan analis diskriminan diantaranya adalah sebagai berikut:

- Sensitif terhadap outlier, rentan terhadap data pencilan yang dapat mempengaruhi hasil analisis.
- Memerlukan banyak data, membutuhkan jumlah data yang cukup besar dalam setiap kelompok untuk hasil yang baik.
- Kurang efektif untuk data non-linier.

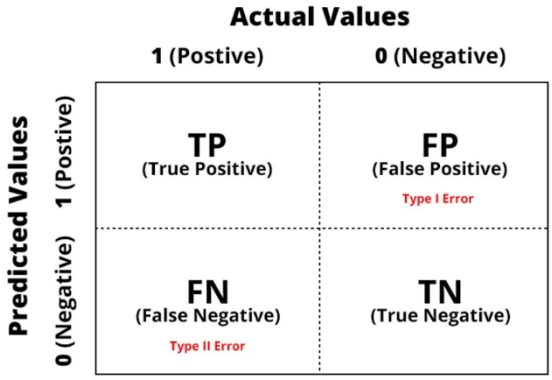
2.1 Evaluasi Model

Evaluasi model merupakan tahap krusial yang membantu untuk menilai sejauh mana keberhasilan model dalam memprediksi dengan tepat. Baik dalam konteks *data mining*, *text mining*, maupun *machine learning*, evaluasi model sangat penting untuk memastikan keefektifan dan keakuratan prediksi. Dengan evaluasi model, dapat memberikan gambaran tentang seberapa baik model dalam menjalankan tugas yang diinginkam. Terdapat beberapa metode dan ukuran evaluasi yang umum digunakan, seperti akurasi (*accuracy*), presisi (*precision*), sensitivitas (*recall*), dan *F1-score*. Dengan memahami evaluasi model, kita dapat mengidentifikasi kelebihan dan kelemahan dari model yang telah dibangun. Sehingga dapat melakukan perbaikan yang diperlukan agar model memberikan hasil prediksi yang lebih optimal.

2.6.1 *Confusion Matrix*

Confusion matrix atau *error matrix* merupakan alat yang digunakan untuk mengevaluasi kinerja model klasifikasi pada data uji yang hasil sebenarnya

sudah diketahui. Tujuan utama dari *confusion matrix* adalah untuk memvisualisasikan dan menganalisis hasil prediksi yang dibuat oleh model, sehingga memudahkan dalam memahami kelebihan dan kekurangan model dalam mengklasifikasikan data. Matriks ini terdiri dari empat komponen utama yang tersusun dalam bentuk tabel sederhana.



Gambar 2.1 *Confusion Matrix*

Berikut komponen-komponen dalam *confusion matrix*:

1. **True Positive (TP)**: Kasus di mana model memprediksi data kelas positif dan benar (prediksi positif, kenyataan positif).
2. **True Negative (TN)**: Kasus di mana model memprediksi data kelas negatif dan benar (prediksi negatif, kenyataan negatif).
3. **False Positive (FP)**: Kasus di mana model memprediksi data kelas positif namun salah (prediksi positif, kenyataan negatif). Ini juga dikenal sebagai "Type I error".
4. **False Negative (FN)**: Kasus di mana model memprediksi data kelas negatif namun salah (prediksi negatif, kenyataan positif). Ini juga dikenal sebagai "Type II error".

Penting untuk memahami posisi setiap elemen dalam *confusion matrix* dengan tepat, karena hal ini akan sangat membantu dalam proses evaluasi model serta meningkatkan pemahaman mengenai kinerja model dalam memprediksi hasil yang diinginkan.

2.6.2 Akurasi (*Accuracy*)

Akurasi adalah metrik evaluasi yang menilai seberapa baik model membuat prediksi yang benar dari keseluruhan prediksi yang dihasilkan. Dalam klasifikasi, akurasi menunjukkan seberapa sering model memprediksi kelas

yang tepat, baik positif maupun negatif. Hal ini berarti akurasi adalah rasio prediksi yang benar (baik positif maupun negatif) terhadap total data yang dinilai. Nilai akurasi (*accuracy*) dapat diperoleh dengan persamaan berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Dengan mengetahui akurasi, kita dapat menilai sejauh mana model berhasil dalam melakukan klasifikasi. Namun, perlu diingat bahwa akurasi mungkin bukan metrik terbaik, terutama ketika data tidak seimbang atau ketika ada perbedaan biaya untuk kesalahan jenis yang berbeda (seperti *False Positive* dan *False Negative*). Oleh karena itu, penting untuk mempertimbangkan metrik evaluasi lain dalam mengukur kinerja model.

2.6.3 Presisi (*Precision*)

Presisi adalah metrik evaluasi yang mengukur seberapa baik model membuat prediksi yang benar untuk kelas positif dari total prediksi positif yang dilakukan. Dalam konteks klasifikasi, presisi memberikan gambaran mengenai seberapa sering model memprediksi kelas positif dengan benar, di antara semua prediksi positif yang dibuat oleh model. Nilai presisi (*precision*) dapat diperoleh dengan persamaan berikut:

$$Precision = \frac{TP}{TP + FP}$$

Dengan mengetahui presisi, kita dapat menilai sejauh mana model berhasil dalam melakukan klasifikasi yang lebih fokus pada kelas positif dan mengurangi kesalahan jenis *False Positive*.

2.6.4 Sensitivitas (*Recall*)

Sensitivitas (*recall*) merupakan evaluasi metrik yang menggambarkan seberapa efektif model dalam mengenali kelas positif secara tepat. Ini menghitung rasio prediksi yang benar positif dibandingkan dengan keseluruhan data yang sebenarnya positif. Nilai sensitivitas (*recall*) dapat diperoleh dengan persamaan berikut:

$$Recall = \frac{TP}{TP + FN}$$

Kelebihan *recall* adalah fokus pada mengurangi kesalahan *False Negative*, sehingga dapat dipastikan bahwa sebanyak mungkin *review* positif diidentifikasi dengan benar.

2.6.5 *F1-Score*

F1-Score adalah evaluasi metrik yang menggambarkan sejauh mana keseimbangan antara Presisi (*precision*) dan Sensitivitas (*recall*). *F1-Score* memberikan informasi seberapa baik model mampu menggabungkan kemampuan Presisi dan Sensitivitas yang penting untuk menilai keakuratan klasifikasi data secara akurat. Kelebihan *F1-Scor*e adalah mempertimbangkan kedua aspek kinerja model (Presisi dan Sensitivitas) dalam satu angka, sehingga bisa mendapatkan gambaran yang lebih lengkap tentang kinerja model. *F1-Score* dapat diperoleh dengan persamaan berikut:

$$F1\ Score = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

Nilai terbaik *F1-Score* adalah 1.0 dan nilai terburuknya adalah 0. Secara representasi, jika *F1-Score* memiliki skor yang baik mengindikasikan bahwa model klasifikasi memiliki *precision* dan *recall* yang baik. Secara umum, nilai *F1-Score* dikatakan tinggi jika berada dalam rentang berikut:

- 0.9 - 1.0: Sangat tinggi. Model memiliki performa sangat baik dengan tingkat *precision* dan *recall* yang sangat tinggi.
- 0.8 - 0.9: Tinggi. Model berkinerja baik, meskipun masih ada kemungkinan beberapa kesalahan klasifikasi.
- 0.7 - 0.8: Cukup tinggi. Model menunjukkan performa yang baik, namun masih ada ruang untuk perbaikan.
- 0.6 - 0.7: Sedang. Model memiliki performa menengah dan perbaikan signifikan mungkin diperlukan.
- 0.5 - 0.6: Rendah. Model menunjukkan performa kurang baik dengan banyak kesalahan klasifikasi.
- < 0.5: Sangat rendah. Model memiliki performa yang buruk dan tidak dapat diandalkan.

BAB III
METODOLOGI PENELITIAN

3.1 Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan jenis penelitian kuantitatif dengan pendekatan deduktif dan eksperimental. Penelitian kuantitatif dipilih karena menggunakan data numerik dan statistik untuk menganalisis dan membandingkan akurasi dua metode klasifikasi, yaitu regresi logistik biner dan analisis diskriminan. Sedangkan pendekatan eksperimental dipilih karena penelitian ini melibatkan manipulasi variabel independen (metode klasifikasi) untuk melihat efeknya pada variabel dependen (akurasi klasifikasi). Selanjutnya, Pendekatan deduktif juga dipilih karena penelitian ini dimulai dengan hipotesis yang jelas tentang mana yang lebih akurat antara regresi logistik biner dan analisis diskriminan. Hipotesis ini kemudian akan diuji dengan data dan analisis statistik. Peneliti akan menerapkan kedua metode klasifikasi pada *dataset* yang sama dan membandingkan tingkat akurasinya.

3.2 Data dan Sumber Data

Data penelitian ini diperoleh dari Kaggle, platform yang menyediakan dataset publik untuk berbagai analisis. Dataset yang digunakan dalam penelitian ini berisi informasi tentang pinjaman bank, termasuk variabel-variabel berikut:

1. ID: ID Pelanggan
2. Age: Usia Pelanggan
3. Experience: Pengalaman Kerja Pelanggan
4. Income: Pendapatan Pelanggan
5. ZipCode: Kode Pos Tempat Tinggal Pelanggan
6. Family: Jumlah Anggota Keluarga Pelanggan
7. CCAvg: Rata-rata Penggunaan Kartu Kredit
8. Education: Pendidikan Pelanggan
9. Mortgage: Hipotek yang Diambil atau Tidak oleh Pelanggan
10. Securities Account: Memiliki (1) atau Tidak Memiliki (0) Rekening Efek
11. CD Account: Memiliki (1) atau Tidak Memiliki (0) Rekening Deposit Bersertifikat (CD)
12. Online: Menggunakan (1) atau Tidak Menggunakan (0) Layanan Perbankan Online
13. Credit Card: Memiliki (1) atau Tidak Memiliki (0) Kartu Kredit

14. Personal Loan: Pinjaman Pribadi (0 = Tidak diberikan pinjaman pribadi, 1 = Diberikan pinjaman pribadi)

3.3 Alat dan Teknik Analisis

Penelitian ini menggunakan *software* Rstudio sebagai alat analisis. Adapun teknik analisis yang digunakan adalah berupa teknik klasifikasi data menggunakan dua metode, yaitu regresi logistik biner dan analisis diskriminan. Adapun setelah tahap *preprocessing* data, data akan dibagi menjadi dua yaitu 70% untuk data *training* dan 30% untuk data *testing*. Hasil analisis akhirnya berupa perbandingan antara kedua metode tersebut menggunakan hasil evaluasi model akhir dan nilai akurasinya.

BAB IV
ANALISIS DAN PEMBAHASAN

4.1 Preprocessing Data

1. Memanggil data yang akan digunakan dalam pengujian.

```
> data_bankloan <- read.csv("C:/Lessons/SEM 6/bankloan.csv", sep = ',')
> head(data_bankloan)
  ID Age Experience Income ZIP.Code Family CCAvg Education Mortgage Personal.Loan Securities.Account CD.Account Online CreditCard
1  1  25          1     49    91107      4   1.6         1        0            0                 1          0      0          0
2  2  45         19     34    90089      3   1.5         1        0            0                 1          0      0          0
3  3  39         15     11    94720      1   1.0         1        0            0                 0          0      0          0
4  4  35          9    100    94112      1   2.7         2        0            0                 0          0      0          0
5  5  35          8     45    91330      4   1.0         2        0            0                 0          0      0          1
6  6  37         13     29    92121      4   0.4         2      155            0                 0          0      1          0
```

2. Menghapus kolom yang tidak akan digunakan dalam pengujian, yaitu kolom "ID".

```
> data_bankloan <- data_bankloan[ , !(names(data_bankloan) %in% c("ID"))]
> head(data_bankloan)
  Age Experience Income ZIP.Code Family CCAvg Education Mortgage Personal.Loan Securities.Account CD.Account Online CreditCard
1  25          1     49    91107      4   1.6         1        0            0                 1          0      0          0
2  45         19     34    90089      3   1.5         1        0            0                 1          0      0          0
3  39         15     11    94720      1   1.0         1        0            0                 0          0      0          0
4  35          9    100    94112      1   2.7         2        0            0                 0          0      0          0
5  35          8     45    91330      4   1.0         2        0            0                 0          0      0          1
6  37         13     29    92121      4   0.4         2      155            0                 0          0      1          0
```

3. Memindahkan kolom Personal.Loan (variabel y) ke kolom terakhir agar memudahkan pengujian yang akan dilakukan.

```
> data_bankloan <- data_bankloan[, c(setdiff(names(data_bankloan), "Personal.Loan"), "Personal.Loan")]
> head(data_bankloan)
  Age Experience Income ZIP.Code Family CCAvg Education Mortgage Securities.Account CD.Account Online CreditCard Personal.Loan
1  25          1     49    91107      4   1.6         1        0                 1          0      0          0            0
2  45         19     34    90089      3   1.5         1        0                 1          0      0          0            0
3  39         15     11    94720      1   1.0         1        0                 0          0      0          0            0
4  35          9    100    94112      1   2.7         2        0                 0          0      0          0            0
5  35          8     45    91330      4   1.0         2        0                 0          0      0          1            0
6  37         13     29    92121      4   0.4         2      155                 0          0      1          0            0
```

4. Mengubah variabel Personal.Loan menjadi variabel factor.

```
> # Mengubah Personal.Loan menjadi faktor
> data_bankloan$Personal.Loan <- as.factor(data_bankloan$Personal.Loan)
> str(data_bankloan)
'data.frame':   3821 obs. of  13 variables:
 $ Age               : int  39 35 35 37 53 50 35 65 29 59 ...
 $ Experience        : int  15 9 8 13 27 24 10 39 5 32 ...
 $ Income            : int  11 100 45 29 72 22 81 105 45 40 ...
 $ ZIP.Code          : int  94720 94112 91330 92121 91711 93943 90089 94710 90277 94920 ...
 $ Family            : int  1 1 4 4 2 1 3 4 3 4 ...
 $ CCAvg             : num  1 2.7 1 0.4 1.5 0.3 0.6 2.4 0.1 2.5 ...
 $ Education         : int  1 2 2 2 2 3 2 3 2 2 ...
 $ Mortgage          : int  0 0 0 155 0 0 104 0 0 0 ...
 $ Securities.Account: int  0 0 0 0 0 0 0 0 0 0 ...
 $ CD.Account        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Online            : int  0 0 0 1 1 0 1 0 1 1 ...
 $ CreditCard        : int  0 0 1 0 0 1 0 0 0 0 ...
 $ Personal.Loan     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

5. Mengecek dan menghapus *missing value.*

```
> sum(is.na(data_bankloan))
[1] 0
> # Menghapus baris dengan missing values
> data_bankloan<- na.omit(data_bankloan)
```

Berdasarkan *output*, diperoleh bahwa tidak terdapat *missing* value pada data. Sehingga preprocessing dapat dilanjutkan.

6. Menghapus outlier pada data.

```
> dimensi_data_bankloan <- dim(data_bankloan)
> dimensi_data_bankloan
[1] 5000   13
> # 2. Menghitung IQR dan menghilangkan outlier untuk beberapa kolom
> remove_outliers <- function(data_bankloan, columns) {
+   for (col in columns) {
+     Q1 <- quantile(data_bankloan[[col]], 0.25, na.rm = TRUE)
+     Q3 <- quantile(data_bankloan[[col]], 0.75, na.rm = TRUE)
+     IQR <- Q3 - Q1
+     data_bankloan <- data_bankloan[!(data_bankloan[[col]] < (Q1 - 1.5 * IQR) | data_bankloan[[col]] > (Q3 + 1.5 * IQR)), ]
+   }
+   return(data_bankloan)
+ }
> # Daftar kolom yang ingin dihilangkan outlier
> columns_to_process <- c("Age","Experience", "Income", "ZIP.Code", "Family","CCAvg", "Education", "Mortgage","Securities.Account","CD.Accoun
t","Online", "CreditCard")
> # Menghapus outlier dari kolom yang dipilih
> data_bankloan <- remove_outliers(data_bankloan, columns_to_process)
> dimensi_data_bankloan <- dim(data_bankloan)
> dimensi_data_bankloan
[1] 3821   13
```

Berdasarkan *output*, diperoleh bahwa dimensi data berubah menjadi 3821 baris dengan 13 kolom, yang dimana sebelum melakukan penghapusan pada outlier dimensi data tersebut adalah 5000 baris dengan 13 kolom. Artinya terdapat 1279 data yang merupakan outlier.

7. Menghapus kolom yang memiliki standar deviasi = 0.

```
> sd_values <- apply(data_bankloan[, 1:12], 2, sd)
> sd_values
           Age     Experience         Income        ZIP.Code         Family          CCAvg      Education
     11.5176746     11.5297239     37.3328620   1764.0738653      1.1602153      1.1354812      0.8371514
      Mortgage Personal.Loan Securities.Account     CD.Account         Online
     66.2192321      0.2206286      0.0000000      0.0000000      0.4931996
> # Identifikasi kolom yang memiliki standar deviasi nol
> columns_to_remove_sd <- which(sd_values == 0)
> # Periksa kolom yang akan dihapus
> if(length(columns_to_remove_sd) > 0) {
+   print("Kolom yang akan dihapus karena std = 0:")
+   print(names(data_bankloan)[columns_to_remove_sd])
+ } else {
+   print("Tidak ada kolom dengan standar deviasi nol.")
+ }
[1] "Kolom yang akan dihapus karena std = 0:"
[1] "Securities.Account" "CD.Account"
> # Hapus kolom yang memiliki standar deviasi nol
> if(length(columns_to_remove_sd) > 0) {
+   data_bankloan_clean <- data_bankloan[, -columns_to_remove_sd]
+ } else {
+   data_bankloan_clean <- data_bankloan
+ }
> dim(data_bankloan_clean)
[1] 3821   11
> head(data_bankloan_clean)
  Age Experience Income ZIP.Code Family CCAvg Education Mortgage Personal.Loan Online CreditCard
3  39         15     11    94720      1   1.0         1        0             0      0          0
4  35          9    100    94112      1   2.7         2        0             0      0          0
5  35          8     45    91330      4   1.0         2        0             0      0          1
6  37         13     29    92121      4   0.4         2      155             0      1          0
7  53         27     72    91711      2   1.5         2        0             0      1          0
8  50         24     22    93943      1   0.3         3        0             0      0          1
```

Berdasarkan *output* diperoleh bahwa, dari 12 variabel independen terdapat 2 variabel yang memiliki nilai standar deviasi = 0, yaitu *Securities. Account* dan *CD. Account.* Sehingga dua variabel tersebut dihapus dan tidak digunakan dalam pengujian yang akan dilakukan.

8. Menghapus kolom yang sangat berkorelasi.

```
> y=data_bankloan_clean[ , 11]
> x=data_bankloan_clean[ , 1:10]
> # Menghitung matriks korelasi
> cor_matrix <- cor(x)
> head(cor_matrix)
                    Age  Experience      Income    ZIP.Code      Family       CCAvg    Education     Mortgage      Online   CreditCard
Age         1.00000000  0.99414111 -0.03639261 -0.03482080 -0.06184317 -0.03085522  0.036030869 -0.010452630  0.02214679  0.007702265
Experience  0.99414111  1.00000000 -0.02976805 -0.03493734 -0.06782181 -0.03471910  0.008830572 -0.008904621  0.02066238  0.008386652
Income     -0.03639261 -0.02976805  1.00000000 -0.03304986 -0.13728998  0.51188089 -0.169939660 -0.074081586 -0.01612459 -0.022259300
ZIP.Code   -0.03482080 -0.03493734 -0.03304986  1.00000000  0.02128750 -0.02151679 -0.004859066  0.005972148  0.03299733  0.014334960
Family     -0.06184317 -0.06782181 -0.13728998  0.02128750  1.00000000 -0.04345854  0.029658412  0.032464909  0.01902453  0.007242904
CCAvg      -0.03085522 -0.03471910  0.51188089 -0.02151679 -0.04345854  1.00000000 -0.077356346 -0.027033212 -0.03249293 -0.015946272
> # Identifikasi kolom yang sangat berkorelasi (korelasi > 0.99)
> columns_to_remove_cor <- findCorrelation(cor_matrix, cutoff = 0.99)
> # Periksa kolom yang akan dihapus
> if(length(columns_to_remove_cor) > 0) {
+   print("Kolom yang akan dihapus karena korelasi > 0.99:")
+   print(names(data_bankloan)[columns_to_remove_cor])
+ } else {
+   print("Tidak ada kolom yang sangat berkorelasi")
+ }
[1] "Kolom yang akan dihapus karena korelasi > 0.99:"
[1] "Age"
> # Hapus kolom yang sangat berkorelasi
> if(length(columns_to_remove_cor) > 0) {
+   data_bankloan_clean <- data_bankloan_clean[, -columns_to_remove_cor]
+ } else {
+   data_bankloan_clean <- data_bankloan_clean
+ }
```

Berdasarkan *output* diatas, diperoleh informasi bahwa terdapat satu variabel yang berkorelasi tinggi, yaitu variabel *Age.* Sehingga variabel *Age* dihapus dan tidak digunakan dalam pengujian yang akan dilakukan.

9. Menghapus kolom yang memiliki nilai IQR = 0.

```
> iqr_values <- apply(data_bankloan_clean[, 1:9], 2, IQR)
> iqr_values
Experience     Income   ZIP.Code     Family      CCAvg  Education   Mortgage     Online CreditCard
      20.0       49.0     2668.0        3.0        1.6        2.0       81.0        1.0        1.0
> # Identifikasi kolom yang memiliki IQR 0
> columns_to_remove_iqr <- which(iqr_values == 0)
> # Periksa kolom yang akan dihapus karena IQR = 0
> if(length(columns_to_remove_iqr) > 0) {
+   print("Kolom yang akan dihapus karena IQR = 0:")
+   print(names(data_bankloan_clean)[columns_to_remove_iqr])
+ } else {
+   print("Tidak ada kolom dengan IQR nol.")
+ }
[1] "Tidak ada kolom dengan IQR nol."
> # Hapus kolom yang memiliki IQR nol jika ada
> if(length(columns_to_remove_iqr) > 0) {
+   data_bankloan_clean <- data_bankloan_clean[, -columns_to_remove_iqr]
+ }
> head(data_bankloan_clean)
  Experience Income ZIP.Code Family CCAvg Education Mortgage Online CreditCard Personal.Loan
3         15     11    94720      1   1.0         1        0      0          0             0
4          9    100    94112      1   2.7         2        0      0          0             0
5          8     45    91330      4   1.0         2        0      0          1             0
6         13     29    92121      4   0.4         2      155      1          0             0
7         27     72    91711      2   1.5         2        0      1          0             0
8         24     22    93943      1   0.3         3        0      0          1             0
```

Berdasarkan *output* diatas, diperoleh informasi bahwa pada tahapan terkahir preprocessing data, yaitu menghapus kolom yang memiliki nilai IQR=0. Hasil *output* menunjukkan bahwa tidak terdapat variabel yang memiliki nilai IQR=0. Sehingga, diperoleh variabel independen akhir yang akan digunakan dalam pengujian, diantaranya adalah *Experience, Income, ZIP. Code, Family, CCAvg, Education, Mortgage, Online, Credit Card,* dan *Personal.Loan* sebagai variabel dependen.

10. Membagi data, dengan data training 70% dan data testing 30%.

```
> data_bankloan<-data_bankloan_clean
> set.seed(123)
> #Membagi data menjadi training dan testing
> indeks <- createDataPartition(data_bankloan$Personal.Loan, p = 0.7, list = FALSE)
> data_training <- data_bankloan[indeks, ]
> data_testing <- data_bankloan[-indeks, ]
> dim(data_training)
[1] 2676   10
> dim(data_testing)
[1] 1145   10
> dim(data_bankloan)
[1] 3821   10
```

Data yang telah dilakukan *cleaning* atau telah melewati tahapan preprocessing data, dibagi menjadi data training dan data testing dengan proporsi 70% data training dan 30% data testing. Kemudian, data training akan digunakan dalam analisis atau pengujian sedangkan data testing akan digunakan dalam menghitung evaluasi prediksi.

4.2 Regresi Logistik Biner

Setelah dilakukan tahap *preprocessing* data, didapatkan hanya sembilan variabel independen yang akan dilakukan pengujian, yaitu variabel *Experience, Income, ZIP.Code, Family, CCAvg, Education, Mortgage, Online, CreditCard.*

4.2.1 Uji Rasio Likelihood

- Hipotesis

  $H_0: \beta_1 = \beta_{\dots} = \beta_9 = 0$ (Secara bersama-sama variabel bebas tidak memengaruhi model)

  H_1: Salah satu dari $\beta_k \neq 0$ dengan k=1,2,..,9 (Secara bersama-sama variabel bebas memengaruhi model)

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik Uji

  Berdasarkan *output* hasil model logistik pada Rstudio, didapatkan nilai:

  $G = -2 \ln \left(\frac{Likelihood\ tanpa\ variabel\ bebas}{Likelihood\ dengan\ variabel\ bebas}\right) = 1098.559 - 483.044$

  $= 615.515$

  Dan nilai p-value = 9.78709233437347e-127

  $\chi^2_{(\alpha,df)} = \chi^2_{(0.05,9)} = 16.91898$

- Kriteria Uji

  Tolak H_0 jika nilai G > $X^2_{(\alpha,p)}$ atau p-value < α

- Keputusan

  H_0 ditolak karena nilai G (615.515) > (16.91898) $X^2_{(\alpha,9)}$ dan p-value $(9.78709233437347\text{e-}127) < \alpha\ (0.05)$

- Kesimpulan

  Pada taraf signifikansi $\alpha = 5\%$, H_0 ditolak. Maka dapat disimpulkan bahwa secara bersama-sama variabel mempengaruhi model.

4.2.2 Uji Goodness of Fit

- Hipotesis

  H_0: Model sesuai (observasi dan prediksi tidak berbeda)

  H_1: Model tidak sesuai (observasi dan prediksi berbeda)

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik Uji

  Berdasarkan *output* Hosmer-Lemeshow Test, didapatkan nilai:

  C = $\sum_{k=1}^{g} \frac{(O_k - n\pi_k)^2}{n_k \pi_k (1-\pi_k)} = 9.3532$

  Dan nilai p-value = 0.4054

  $\chi^2_{(\alpha,g-2)} = \chi^2_{(0.05,11-7)} = \chi^2_{(0.05,9)} = 16.91898$

- Kriteria Uji

  Tolak H_0 jika nilai C > $X^2_{(\alpha,g-2)}$ atau p-value < α

- Keputusan

  H_0 gagal ditolak karena nilai C $(9.3532) < (16.91898)$ $X^2_{(\alpha,9)}$ dan p-value $(0.4054) > \alpha\ (0.05)$

- Kesimpulan

  Pada taraf signifikansi $\alpha = 5\%$, H_0 gagal ditolak. Maka dapat disimpulkan bahwa model sesuai atau tidak ada perbedaan antara observasi dan prediksi.

4.2.3 Uji Wald

- Hipotesis

  H_0: $\beta_j = 0$, j=1,2,..,9 (Variabel bebas tidak signifikan terhadap model)

  H_1: $\beta_j \neq 0$, j=1,2,..,9 (Variabel bebas signifikan terhadap model)

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik Uji

  Wj = $\left(\frac{\widehat{\beta_J}}{Se\widehat{\beta_J}}\right)^2$

  - Nilai Wald Intercept = 4.278602; nilai p-value = 0.03859509
  - Nilai Wald Experience = 8.783916e-05; p-value = 0.9925221
  - Nilai Wald Income = 174.8177; nilai p-value = 6.561574e-40
  - Nilai Wald ZIP.Code = 0.07226572; nilai p-value = 0.788066
  - Nilai Wald Family = 20.22398; nilai p-value = 6.888386e-06
  - Nilai Wald CCAvg = 41.90271; nilai p-value = 9.592944e-11
  - Nilai Wald Education = 80.07298; nilai p-value = 3.608328e-19
  - Nilai Wald Mortgage = 0.2242773; nilai p-value = 0.6358
  - Nilai Wald Online = 1.829066; nilai p-value = 0.176238
  - Nilai Wald CreditCard = 5.143948; nilai p-value = 0.02332753

$\chi^2_{(\alpha,df)} = \chi^2_{(0.05,1)} = 3.8415$

- Kriteria Uji

  Tolak H_0 jika $w_j > \chi^2_{(\alpha,1)}$ atau p-value < α

- Keputusan

Variabel	Wald	p-value	Keputusan H0
Intercept	4.278602	0.03859509	Ditolak
Experience	8.783916e-05	0.9925221	Diterima
Income	174.8177	6.561574e-40	Ditolak
ZIP.Code	0.07226572	0.788066	Diterima
Family	20.22398	6.888386e-06	Ditolak
CCAvg	41.90271	9.592944e-11	Ditolak
Education	80.07298	3.608328e-19	Ditolak
Mortgage	0.2242773	0.6358	Diterima
Online	1.829066	0.176238	Diterima
CreditCard	5.143948	0.02332753	Ditolak

- Kesimpulan

  Berdasarkan hasil tabel di atas, pada taraf signifikansi $\boldsymbol{\alpha}$ = 5%, untuk variabel Intercept, *Income, Family, CCAvg, Education, CreditCard*, H_0 ditolak karena nilai $w_j > \chi^2_{(0.05,1)}$ (3.8415) atau p-value < α (0.05) sehinggga variabel tersebut signifikan terhadap model. Sedangkan untuk variabel *Experience, ZIP.Code, Mortgage, Online,* H_0 gagal ditolak karena nilai $w_j \leq \chi^2_{(0.05,1)}$ (3.8415) atau p-value ≥ α (0.05) atau variabel tersebut tidak signifikan terhadap model.

Karena terdapat variabel bebas yang tidak signifikan, maka dilakukan uji hipotesis ulang sebagai berikut:

4.2.4 Uji Rasio Likelihood Variabel Signifikan

- Hipotesis

  $H_0: \beta_1 = \beta_{..} = \beta_5 = 0$ (Secara bersama-sama variabel bebas tidak memengaruhi model)

H_1: Salah satu dari $\beta_k \neq 0$ dengan k=1,2,..,5 (Secara bersama-sama variabel bebas memengaruhi model)

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik Uji

  Berdasarkan *output* hasil model logistik pada Rstudio, didapatkan nilai:

  G = $-2 \ln \left(\frac{Likelihood\ tanpa\ variabel\ bebas}{Likelihood\ dengan\ variabel\ bebas}\right) = 1098.559 - 485.214$

  $= 613.345$

  Dan nilai p-value = 2.64478920899507e-130

  $\chi^2_{(\alpha,df)} = \chi^2_{(0,05,5)} = 11.0705$

- Kriteria Uji

  Tolak H_0 jika nilai G > $X^2_{(\alpha,5)}$ atau p-value < α

- Keputusan

  H_0 ditolak karena nilai G (613.345) > (11.0705) $X^2_{(\alpha,9)}$ dan p-value $(2.64478920899507\text{e-}130) < \alpha\ (0.05)$

- Kesimpulan

  Pada taraf signifikansi $\alpha = 5\%$, H_0 ditolak. Maka dapat disimpulkan bahwa secara bersama-sama variabel mempengaruhi model.

4.2.5 Uji Goodness of Fit Variabel Signifikan

- Hipotesis

  H_0: Model sesuai (observasi dan prediksi tidak berbeda)

  H_1: Model tidak sesuai (observasi dan prediksi berbeda)

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik Uji

  Berdasarkan *output* Hosmer-Lemeshow Test, didapatkan nilai:

  C = $\sum_{k=1}^{g} \frac{(O_k - n\pi_k)^2}{n_k \pi_k (1-\pi_k)} = 10.288$

  Dan nilai p-value = 0.06746

  $\chi^2_{(\alpha,g-2)} = \chi^2_{(0,05,7-2)} = \chi^2_{(0.05,5)} = 11.0705$

- Kriteria Uji

  Tolak H_0 jika nilai C > $X^2_{(\alpha,g-2)}$ atau sig < α

- Keputusan

  H_0 gagal ditolak karena nilai C $(10.288) < (11.0705)\ X^2_{(0.05,5)}$ dan p-value $(0.06746) > \alpha\ (0.05)$

- Kesimpulan

  Pada taraf signifikansi $\alpha = 5\%$, H_0 gagal ditolak. Maka dapat disimpulkan bahwa model sesuai atau tidak ada perbedaan antara observasi dan prediksi.

4.2.6 Uji Wald Variabel Signifikan

- Hipotesis

  H_0: $\beta_j = 0$, j=1,2,..,5 (Variabel bebas tidak signifikan terhadap model)

  H_1: $\beta_j \neq 0$, j=1,2,..,5 (Variabel bebas signifikan terhadap model)

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik Uji

  Wj = $\left(\frac{\widehat{\beta_j}}{Se\widehat{\beta_j}}\right)^2$

  - Nilai Wald Intercept = 229.0689; nilai p-value = 9.514793e-52
  - Nilai Wald Income = 177.5849; nilai p-value = 1.632125e-40
  - Nilai Wald Family = 20.46246; nilai p-value = 6.081242e-06
  - Nilai Wald CCAvg = 42.12378; nilai p-value = 8.567483e-11
  - Nilai Wald Education = 80.96471; nilai p-value = 2.297849e-19
  - Nilai Wald CreditCard = 4.196726; nilai p-value = 0.0405021

  $\chi^2_{(\alpha,df)} = \chi^2_{(0,05,1)} = 3.8415$

- Kriteria Uji

  Tolak H_0 jika $w_j > \chi^2_{(\alpha,db)}$ atau p-value < α

- Keputusan

Variabel	Wald	p-value	Keputusan H0
Intercept	229.0689	9.514793e-52	Ditolak
Income	177.5849	1.632125e-40	Ditolak
Family	20.46246	6.081242e-06	Ditolak
CCAvg	42.12378	8.567483e-11	Ditolak
Education	80.96471	2.297849e-19	Ditolak

CreditCard	4.196726	0.0405021	Ditolak

- Kesimpulan

  Berdasarkan hasil tabel di atas, pada taraf signifikansi $\boldsymbol{\alpha}$ = 5%, semua variabel (Intercept, *Income, Family, CCAvg, Education, CreditCard*), H_0 ditolak karena nilai $w_j > \chi^2_{(0.05,1)}$ (3.8415) atau p-value < α (0.05) sehinggga semua variabel tersebut signifikan terhadap model.

4.2.7 Model Akhir Variabel Signifikan

Setelah dilakukan uji-uji di atas, didapatkan model akhir regresi logistik biner. Adapun model akhir hanya mengikutkan parameter dari variabel-variabel yang signifikan saja.

Variabel: (Intercept) Estimasi Koefisien: -15.1792 Kuadrat Standar Error: 1.005846 Statistik uji Wald: 229.0689 Nilai p-value: 9.514793e-52 Tolak H0: Variabel bebas signifikan terhadap model	Variabel: Income Estimasi Koefisien: 0.064439 Kuadrat Standar Error: 2.338253e-05 Statistik uji Wald: 177.5849 Nilai p-value: 1.632125e-40 Tolak H0: Variabel bebas signifikan terhadap model
Variabel: Family Estimasi Koefisien: 0.5406403 Kuadrat Standar Error: 0.0142843 Statistik uji Wald: 20.46246 Nilai p-value: 6.081242e-06 Tolak H0: Variabel bebas signifikan terhadap model	Variabel: CCAvg Estimasi Koefisien: 0.61929 Kuadrat Standar Error: 0.009104597 Statistik uji Wald: 42.12378 Nilai p-value: 8.567483e-11 Tolak H0: Variabel bebas signifikan terhadap model
Variabel: Education Estimasi Koefisien: 1.63328 Kuadrat Standar Error: 0.03294774 Statistik uji Wald: 80.96471 Nilai p-value: 2.297849e-19 Tolak H0: Variabel bebas signifikan terhadap model	Variabel: CreditCard Estimasi Koefisien: -0.5936488 Kuadrat Standar Error: 0.08397471 Statistik uji Wald: 4.196726 Nilai p-value: 0.0405021 Tolak H0: Variabel bebas signifikan terhadap model

Berdasarkan *output* Uji Wald di Rstudio, didapat model prediksi akhirnya adalah:

$$\pi(x_i) = \frac{e^{-15.1792 + 0.064439\ Income + 0.5406403\ Family + 0.61929\ CCAvg + 1.63328\ Education - 0.5936488\ CreditCard}}{1+e^{-15.1792 + 0.064439\ Income + 0.5406403\ Family + 0.61929\ CCAvg + 1.63328\ Education - 0.5936488\ CreditCard}}$$

Berdasarkan model tersebut akan diidentifikasi prediksi dari masing-masing individu. Pada penelitian kali ini, digunakan ambang batas *(threshold)* = 0.5. Adapun kriteria prediksinya adalah:

- Jika nilai dari model prediksi akhirnya ≥ *threshold* (0.5) : akan menjadi 1 (diberikan pinjaman pribadi)
- Jika nilai dari model prediksi akhirnya < *threshold* (0.5) : akan menjadi 0 (tidak diberikan pinjaman pribadi)

4.2.8 Uji Non-Multikolinearitas

Uji Multikolinieritas dilakukan untuk melihat apakah ada keterkaitan antara hubungan yang sempurna antara variabel-variabel independen. Jika di dalam pengujian ternyata didapatkan sebuah kesimpulan bahwa antara variabel *independent* tersebut saling terikat, maka pengujian tidak dapat dilakukan ke dalam tahapan selanjutnya yang disebabkan oleh tidak dapat ditentukannya koefisien regresi variabel tersebut.

- Hipotesis

  H_0: Tidak terjadi multikolinieritas

  H_1: Terjadi multikolinieritas
- Taraf Signifikansi

  $\alpha = 5\%$
- Statistik uji:

  Berdasarkan *output* diatas diperoleh nilai sebagai berikut:

  VIF (Income) = 1.941874

  VIF (Family) = 1.387365

  VIF (CCAvg) = 1.053960

  VIF (Education) = 1.715213

  VIF (CreditCard) = 1.006583
- Daerah Kritis

  H_0 ditolak jika nilai VIF > 10
- Keputusan

  H_0 gagal ditolak karena nilai

VIF (Income) = 1.941874 < 10

VIF (Family) = 1.387365 < 10

VIF (CCAvg) = 1.053960 < 10

VIF (Education) = 1.715213 < 10

VIF (CreditCard) = 1.006583 < 10

- Kesimpulan

  Pada taraf signifikansi $\alpha = 5\%$, H_0 gagal ditolak karena nilai VIF dari semua variabel independen < 10. maka dapat disimpulkan bahwa tidak terjadi multikolinieritas.

4.2.9 *Confusion* Matriks dan Evaluasi Model

Hasil *confusion matriks* dari pemodelan adalah:

	0	**1**
0	1080	10
1	25	31

- True Negative (TN): 1080 (Jumlah orang yang tidak diberikan pinjaman dan diprediksi benar oleh model)
- False Positive (FP): 10 (Jumlah orang yang tidak diberikan pinjaman namun diprediksi salah oleh model sebagai diberikan pinjaman)
- False Negative (FN): 25 (Jumlah orang yang diberikan pinjaman namun diprediksi salah oleh model sebagai tidak diberikan pinjaman)
- True Positive (TP): 31 (Jumlah orang yang diberikan pinjaman dan diprediksi benar oleh model)

Dengan nilai *precision, recall, F1-score,* dan akurasi adalah sebagai berikut:

Precision: 0.75609756097561

Recall: 0.553571428571429

F1 Score: 0.639175257731959

Accuracy: 96.9458987783595 %

Interpretasi:

Precision: 0.756 (75.6%)

Interpretasi: Dari semua orang yang diprediksi akan mendapatkan pinjaman, 75.6% benar-benar layak mendapatkan pinjaman. Sebaliknya, 24.4% dari orang yang diprediksi layak ternyata tidak layak.

Recall: 0.554 (55.4%)

Interpretasi: Dari semua orang yang sebenarnya layak mendapatkan pinjaman, hanya 55.4% yang berhasil diidentifikasi oleh model. Sebaliknya, 44.6% dari orang yang layak tidak terdeteksi.

F1 Score: 0.639 (63.9%)

Interpretasi: Kombinasi precision dan recall menunjukkan keseimbangan yang cukup, tetapi masih perlu ditingkatkan terutama dalam hal recall.

Akurasi: 96.95%

Interpretasi: Sekitar 96.95% dari semua prediksi (baik diberikan atau tidak diberikan pinjaman) adalah benar. Sebaliknya, 3.05% dari semua prediksi adalah salah.

4.3 Analisis Diskriminan

Setelah dilakukan tahap *preprocessing* data, diperoleh sembilan variabel independen yang akan dilakukan pengujian, yaitu variabel *Experience, Income, ZIP.Code, Family, CCAvg, Education, Mortgage, Online, CreditCard.*

4.3.1 Uji Non-Multikolinieritas

Uji Multikolinieritas dilakukan untuk melihat apakah ada keterkaitan antara hubungan yang sempurna antara variabel-variabel independen. Dengan hasil pengujian hipotesis sebagai berikut:

- Hipotesis

  H_0: Tidak terjadi multikolinieritas

  H_1: Terjadi multikolinieritas
- Taraf Signifikansi

  $\alpha = 5\%$
- Statistik uji:

  Berdasarkan *output* diatas diperoleh nilai sebagai berikut:

  VIF Experience = 1.006973

  VIF Income = 1.773899

  VIF ZIP. Code = 1.005585

  VIF Family = 1.032681

  VIF CCAvg = 1.360088

  VIF Education = 1.037378

  VIF Mortage = 1.009770

  VIF Online = 1.005001

VIF CreditCard = 1.004353

- Daerah Kritis

  H_0 ditolak jika nilai VIF > 10

- Keputusan

  H_0 gagal ditolak karena nilai

  VIF Experience = 1.006973 < 10

  VIF Income = 1.773899 < 10

  VIF ZIP. Code = 1.005585 < 10

  VIF Family = 1.032681 < 10

  VIF CCAvg = 1.360088 < 10

  VIF Education = 1.037378 < 10

  VIF Mortage = 1.009770 < 10

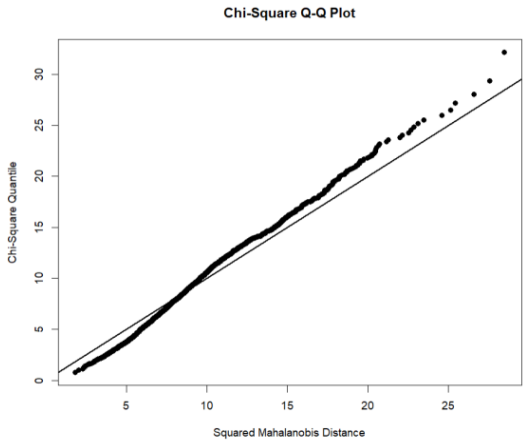
  VIF Online = 1.005001 <10

  VIF CreditCard = 1.004353 < 10

- Kesimpulan

  Pada taraf signifikansi $\alpha = 5\%$, H_0 gagal ditolak karena nilai VIF dari semua variabel independen < 10. Maka dapat disimpulkan bahwa tidak terjadi multikolinieritas.

4.3.2 Normal Univariat dan Multivariat



Gambar 4.1 Chi-Square Q-Q Plot

- Visual

  Berdasarkan chi-square Q-Q Plot diatas, terlihat bahwa plot-plot cenderung membentuk suatu garis lurus. Sehingga dapat disimpulkan bahwa secara visual data berdisrtribusi normal multivariat atau asumsi normalitas multivariat terpenuhi.

- Formal
  - Hipotesis

    H_0: Data berdistribusi normal multivariat

    H_1: Data tidak berdistribusi normal multivariat
  - Taraf Signifikansi

    $\alpha = 5\%$
  - Statistik uji:

    Normal Multivariat

    Berdasarkan *output* diperoleh nilai:

    Mardia Skewness = 2935.0854 dengan p-value = 0.000

    Mardia Kurtosis = -10.92316 dengan p-value = 0.000

    Normal Univariat

    Berdasarkan *output* diperoleh nilai:

    Experience = 24.9096 dengan p-value = <0.001

    Income = 45.1623 dengan p-value = <0.001

    ZIP. Code = 50.9259 dengan p-value = <0.001

    Family = 151. 3774 dengan p-value = <0.001

    CCAvg = 41. 3559 dengan p-value = <0.001

    Education = 239. 6579 dengan p-value = <0.001

    Mortage = 493.9665 dengan p-value = <0.001

    Online = 497. 9487 dengan p-value = <0.001

    CreditCard = 604. 9057 dengan p-value = <0.001
  - Kriteria Uji

    Tolak H_0 jika nilai p-value < α
  - Keputusan

    Normal Multivariat

Test	Statistics	P-value	Keputusan
Mardia Skewness	2935.0854	0.0000	H0 ditolak
Mardia Kurtosis	-10.9232	0.0000	H0 ditolak

    Normal Univariat

Variabel	Statistics	P-value	Keputusan
Experience	24.9096	<0.001	H0 ditolak
Income	45.1623	<0.001	H0 ditolak

ZIP. Code	50.9259	<0.001	H0 ditolak
Family	151. 3774	<0.001	H0 ditolak
CCAvg	41. 3559	<0.001	H0 ditolak
Education	239. 6579	<0.001	H0 ditolak
Mortage	493.9665	<0.001	H0 ditolak
Online	497. 9487	<0.001	H0 ditolak
CreditCard	604. 9057	<0.001	H0 ditolak

- Keputusan

  Pada taraf signifikansi $\alpha = 5\%$, H_0 ditolak untuk semua pengujian baik secara multivariat ataupun univariat. Maka dapat disimpulkan bahwa secara pengujian formal data tidak berdistribusi normal.

4.3.3 Uji Homogenitas

Pengecekan asumsi homogenitas pada data atau pengecekan kesamaan matriks varians kovarians antara dua populasi atau lebih dilakukan dengan Box's M tes dengan hasil sebagai beikut:

- Hipotesis

  H_0 : $\sum_1 = \sum_2 = \ldots \sum_k = \sum$ (matriks kovarians bersifat multivariat heteroskedastisitas)

  H_1 : Minimal ada satu $\sum_i \neq \sum_j$ (matriks kovarians tidak bersifat multivariat heteroskedastisitas)

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik uji:

  Berdasarkan *output* diperoleh nilai sebaga berikut:

  Chi-Square = 171.63

  Df = 45

  P-value = < 2.2 e -16

- Kriteria Uji

  Tolak H_0 jika nilai p-value < α

- Keputusan

  H_0 ditolak karena nilai p-value < α, yaitu (< 2.2 e -16 < 0.05)

- Kesimpulan

Pada taraf signifikansi $\alpha = 5\%$, H_0 ditolak. Maka dapat disimpulkan bahwa matriks kovarians multivariat bersifat heteroskedastisitas. Artinya asumsi homogenitas tidak terpenuhi.

4.3.4 Uji Wilk Lamda

- Hipotesis

  H_0 : Tidak ada perbedaan rata-rata multivariat antara grup

  H_1 : Ada perbedaan rata-rata multivariat antara grup

- Taraf Signifikansi

  $\alpha = 5\%$

- Statistik uji:

  Berdasarkan *output* diperoleh nilai sebaga berikut:

  Wilks Lambda = 0.77139

  Approx F = 87.789

  Num Df = 9

  Den Df = 2666

  P-value = < 2.2 e -16

- Kriteria Uji

  Tolak H_0 jika nilai p-value < α

- Keputusan

  H_0 ditolak karena nilai p-value < α, yaitu < 2.2 e -16 < 0.05

- Kesimpulan

  Pada taraf signifikansi $\alpha = 5\%$, H_0 ditolak. Maka dapat disimpulkan bahwa Ada perbedaan signifikan antara grup (Personal.Loan) dalam hal variabel-variabel dependen gabungan (Experience, Income, ZIP Code, Family, CCAvg, Education, Mortgage, Online, dan CreditCard).

4.3.5 Confusion Matrix dan Evaluasi Model

Hasil *confusion matriks* dari pemodelan adalah:

	0	**1**
0	1074	27
1	31	31

- True Negative (TN): 1074 (Jumlah orang yang tidak diberikan pinjaman dan diprediksi benar oleh model)

- False Positive (FP): 27 (Jumlah orang yang tidak diberikan pinjaman namun diprediksi salah oleh model sebagai diberikan pinjaman)
- False Negative (FN): 31 (Jumlah orang yang diberikan pinjaman namun diprediksi salah oleh model sebagai tidak diberikan pinjaman)
- True Positive (TP): 31 (Jumlah orang yang diberikan pinjaman dan diprediksi benar oleh model)

Dengan nilai *precision, recall, f1 score,* dan akurasi adalah sebagai berikut:

Akurasi = 0.9650655

Precision = 0.5344828

Recall = 0.7045455

F-1 Score = 0.6078431

Interpretasi:

Precision: 0.534 (53.4%)

Interpretasi: Dari semua orang yang diprediksi akan mendapatkan pinjaman, 53.4% benar-benar layak mendapatkan pinjaman. Sebaliknya, 46.6% dari orang yang diprediksi layak ternyata tidak layak.

Recall: 0.705 (70.5%)

Interpretasi: Dari semua orang yang sebenarnya layak mendapatkan pinjaman, hanya 70.5% yang berhasil diidentifikasi oleh model. Sebaliknya, 29.5% dari orang yang layak tidak terdeteksi.

F1 Score: 0.608 (60.8%)

Interpretasi: Kombinasi precision dan recall menunjukkan keseimbangan yang cukup, tetapi masih perlu ditingkatkan terutama dalam hal recall.

Akurasi: 96.5%

Interpretasi: Sekitar 96.5% dari semua prediksi (baik diberikan atau tidak diberikan pinjaman) adalah benar. Sebaliknya, 3.5% dari semua prediksi adalah salah.

4.4 Perbandingan Hasil Regresi Logistik Biner dan Analisis Diskriminan

Dalam analisis ini, perbandingan hasil dari regresi logistik biner dan analisis diskriminan berdasarkan uji asumsi yang telah dilakukan. Berikut merupakan hasil uji yang diperoleh.

Uji	Statistik Uji	Keputusan	Kesimpulan
Uji Rasio Likelihood	G = 613.345, p-value = 2.64e-130	Tolak H0 karena G > χ^2(0.05,5) dan p-value < α	Variabel bebas bersama-sama memengaruhi model.
Uji Goodness of Fit	G = 10.288, p-value = 0.06746	Gagal tolak H0 karena C < χ^2(0.05,5) dan p-value > α	Model sesuai atau tidak ada perbedaan signifikan antara observasi dan prediksi.
Uji Wald	Wj, p-value untuk masing-masing variable	Tolak H0 karena Wj > χ^2(0.05,1) atau p-value < α	Variable Income, Family, CCAvg, Education, CreditCard signifikan terhadap model.
Uji Non-Multikolinieritas	VIF (Income) = 1.941874, VIF (Family) = 1.387365, VIF (CCAvg) = 1.053960, VIF (Education) = 1.715213, VIF (CreditCard) = 1.006583	H0 gagal ditolak karena semua VIF < 10	Tidak terjadi multikolinieritas antara variable *independent* .

Tabel 4.1 Hasil Uji Asumsi Regresi Logistik Biner

Berdasarkan hasil uji yang dilakukan, model regresi logistik biner ini memenuhi beberapa asumsi penting yang diperlukan untuk interpretasi dan penggunaannya.

1. Tidak ada multikolinieritas: Semua nilai Variance Inflation Factor (VIF) untuk variabel independen jauh di bawah ambang batas yang umumnya diterima (VIF < 10). Ini menunjukkan bahwa tidak ada masalah multikolinieritas antara variabel independen.
2. Signifikansi variabel: Variabel yang signifikan terhadap model dipilih berdasarkan uji Wald, di mana variabel yang memiliki pengaruh yang signifikan terhadap respons telah dipertahankan dalam model akhir.
3. Uji Rasio Likelihood: Terdapat bukti bahwa sekumpulan variabel independen secara bersama-sama mempengaruhi model, karena nilai G (statistik uji rasio likelihood) jauh melebihi nilai kritis χ^2 dengan p-value yang sangat kecil.
4. Uji Goodness of Fit: Model tersebut secara keseluruhan sesuai dengan data yang diamati, karena uji Hosmer-Lemeshow menunjukkan bahwa tidak ada perbedaan signifikan antara nilai yang diamati dan nilai yang diprediksi oleh model.

Berdasarkan hasil uji yang dilakukan, model regresi logistik biner tersebut memenuhi sebagian besar asumsi yang diperlukan.

Asumsi	Hasil Uji	Kesimpulan
Uji Non-Multikolinieritas	VIF (Experience) = 1.006973 ; VIF (Income) = 1.773899 ; VIF (ZIP. Code) = 1.005585 ; VIF (Family) = 1.032681 ; VIF (CCAvg) = 1.360088 ; VIF (Education) = 1.037378 ; VIF (Mortgage) = 1.009770 ; VIF (Online) = 1.005001 ; VIF (CreditCard) = 1.004353	Tidak terjadi multikolinieritas karena semua nilai VIF < 10.
Normalitas Multivariat	Mardia Skewness = 2935.0854, p-value <	Data tidak berdistribusi normal multivariat.

	0.001 Mardia Kurtosis = -10.92316, p-value < 0.001	
Normalitas Univariat	Untuk semua variabel independen (Experience, Income, ZIP. Code, Family, CCAvg, Education, Mortgage, Online, CreditCard), p-value < 0.001.	Data tidak berdistribusi normal univariat untuk semua variabel.
Homogenitas Varians	Chi-Square = 171.63, p-value < 0.05	Matriks kovarians multivariat bersifat heteroskedastisitas.
Kesamaan Rata-rata Multivariat	Wilks Lambda = 0.77139, p-value < 0.05	Ada perbedaan signifikan antara grup dalam variabel dependen.

Tabel 4.2 Hasil Uji Asumsi Analisis Diskriminan

1. Non-Multikolinieritas**:** Asumsi terpenuhi karena semua nilai VIF < 10.
2. Normalitas Multivariat**:** Asumsi tidak terpenuhi karena data tidak berdistribusi normal multivariat.
3. Normalitas Univariat**:** Asumsi tidak terpenuhi karena data tidak berdistribusi normal univariat untuk semua variabel.
4. Homogenitas Varians**:** Asumsi tidak terpenuhi karena matriks kovarians multivariat bersifat heteroskedastisitas.
5. Kesamaan Rata-rata Multivariat: Asumsi tidak terpenuhi karena ada perbedaan signifikan antara grup dalam variabel dependen.

Regresi logistik biner pada kasus ini cocok digunakan karena mampu menangani variasi dalam distribusi variabel dan memberikan estimasi probabilitas. Analisis diskriminan mungkin lebih tepat jika variabel-variabel terkait memiliki distribusi normal yang jelas dan tujuan utama adalah pemisahan kelompok. Dalam konteks ini, regresi logistik biner memberikan model yang sesuai dengan variabel-variabel yang signifikan,

tanpa masalah multikolinieritas, dan kesesuaian model yang baik berdasarkan uji Goodness of Fit.

Berdasarkan analisis data yang telah dilakukan maka perbandingan ketepatan klasifikasi kelayakan pinjaman antara analisis regresi logistik biner dan analisis diskriminan diberikan pada Tabel 4.3

Pengukuran	Precision	Recall	F1-Score	Accuracy
Regresi logistik biner	75.6%	55.4%	63.9%	96.95%
Analisis Diskriminan	53.4%	70.5%	60.8%	96.5%

Tabel 4.3 Perbandingan Model

Dari Tabel 4.1 dapat dilihat bahwa terdapat empat metrik evaluasi kinerja model untuk dua jenis model Regresi logistik biner dan Analisis Diskriminan.

1. *Precision*: Regresi logistik biner memiliki precision sebesar 75.6%, sedangkan analisis diskriminan memiliki precision 53.4%. Ini berarti dari semua prediksi positif yang dilakukan oleh model Regresi logistik biner, sekitar 75.6% benar-benar positif, sedangkan untuk Analisis Diskriminan, sekitar 53.4% benar-benar positif.
2. *Recall*: Recall mengukur seberapa banyak dari semua kasus positif yang sebenarnya di dataset, model berhasil memprediksi dengan benar. Analisis diskriminan memiliki recall lebih tinggi (70.5%) dibandingkan regresi logistik biner (55.4%). Ini menunjukkan bahwa analisis diskriminan lebih baik dalam mengidentifikasi kasus positif sebenarnya dalam dataset.
3. *F1-Score*: Regresi logistik biner memiliki F1-score sebesar 63.9%, sedangkan analisis diskriminan memiliki F1-score 60.8%. Meskipun recall analisis iskriminan lebih tinggi, F1-score regresi logistik biner lebih tinggi karena kombinasi yang lebih baik antara precision dan recall.
4. *Accuracy*: Kedua model memiliki tingkat akurasi yang tinggi, regresi logistik biner mencapai 96.95%, sedangkan analisis diskriminan mencapai 96.5%. Namun, perlu diingat bahwa tingkat akurasi dapat menyimpang jika dataset tidak seimbang dalam kelas-kelas yang diprediksi.

BAB V

KESIMPULAN

Berdasarkan hasil uji yang dilakukan, regresi logistik biner menunjukkan bahwa model ini memenuhi sebagian besar asumsi yang penting untuk interpretasi yang tepat. Tidak terdapat masalah multikolinieritas antara variabel independen, variabel-variabel yang signifikan terhadap model telah diidentifikasi melalui uji Wald dan terdapat bukti bahwa sekumpulan variabel independen secara bersama-sama mempengaruhi respons model berdasarkan uji Rasio Likelihood. Selain itu, uji Goodness of Fit menunjukkan bahwa model tersebut sesuai dengan data yang diamati. Meskipun analisis diskriminan menunjukkan bahwa beberapa asumsi, seperti normalitas multivariat dan homogenitas varian, tidak terpenuhi, regresi logistik biner tetap merupakan pilihan yang tepat dalam konteks ini untuk memodelkan variabel-variabel yang signifikan dengan akurat tanpa mengalami kendala dari masalah-masalah tersebut.

Selain itu, berdasarkan hasil analisis perbandingan antara regresi logistik biner dan analisis diskriminan dalam klasifikasi kelayakan pinjaman pada data *bankloan* menggunakan metrik evaluasi, dapat disimpulkan bahwa regresi logistik biner menjadi pilihan model yang lebih unggul dibandingkan analisis diskriminan. Dengan *F1-Score* yang mencapai 63.9%, regresi logistik biner menunjukkan kemampuan yang lebih baik dalam menjaga keseimbangan antara *presicion* dan *recall* dibandingkan analisis diskriminan yang hanya mencapai 60.8%. Selain itu, tingkat akurasi regresi logistik biner yang mencapai 96.95% juga sedikit lebih tinggi dibandingkan dengan analisis diskriminan yang mencatat 96.5%. Meskipun *recall* analisis diskriminan lebih tinggi, *precision* yang signifikan lebih tinggi dari regresi logistik biner (75.6% versus 53.4%) menandakan bahwa regresi logistik biner cenderung memberikan prediksi yang lebih tepat untuk kelas positif. Keputusan untuk memilih regresi logistik biner sebagai model yang lebih akurat sangat didukung oleh kinerja metrik evaluasi yang lebih baik secara keseluruhan.

Secara keseluruhan, berdasarkan analisis yang dilakukan terhadap model regresi logistik biner dan analisis diskriminan dalam konteks klasifikasi kelayakan pinjaman pada dataset *bankloan*, regresi logistik biner menunjukkan keunggulan yang signifikan. Model regresi logistik biner berhasil memenuhi sebagian besar asumsi dan metrik evaluasi yang penting untuk interpretasi yang tepat.

DAFTAR PUSTAKA

Arif Suhendra, Muhammad., Ispriyanti, Dwi., dan Sudarno. 2020. Ketepatan Klasifikasi Pemberian Kartu Keluarga Sejahtera di Kota Semarang Menggunakan Metode Regresi Logistik Biner dan Metode CHAID. *Jurnal Gaussian* Vol. 9, No.1, Hal: 64-74.

Sutrisno dan Wulandari, Dewi. 2018. *Multivariate Analysis of Variance* (MANOVA) untuk Memperkaya Hasil Penelitian Pendidikan. *Jurnal Aksioma* Vol. 9, No.1.

Wikipedia. 2024. *Analisis Diskriminan Linier.* Tersedia: https://en.wikipedia.org/wiki/Linear_discriminant_analysis (diakses pada tanggal 17 Juni 2024).

RPubs by RStudio. 2020. *Discriminant Analysis Manual.* Tersedia: https://rpubs.com/nadhifanhf/discriminant (diakses pada tanggal 17 Juni 2024).

RPubs by RStudio. 2022. *Penerapan Multivariate Analysis of Variance (MANOVA) pada Pengaruh Covid-19 terhadap Intensitas Pembelian Mahasiswa pada Ecommerce (Studi pada Mahasiswa Statistika Universitas Brawijaya Angkatan 2019).* Tersedia: https://rpubs.com/ahmadulfi/miniprojectkomstatg (diakses pada tanggal 17 Juni 2024).

LAMPIRAN

Lampiran 1. Dataset

ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card
1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
3	39	15	11	94720	1	1	1	0	0	0	0	0	0
4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
5	35	8	45	91330	4	1	2	0	0	0	0	0	1
6	37	13	29	92121	4	0.4	2	155	0	0	0	1	0
7	53	27	72	91711	2	1.5	2	0	0	0	0	1	0
8	50	24	22	93943	1	0.3	3	0	0	0	0	0	1
9	35	10	81	90089	3	0.6	2	104	0	0	0	1	0
10	34	9	180	93023	1	8.9	3	0	1	0	0	0	0
11	65	39	105	94710	4	2.4	3	0	0	0	0	0	0
12	29	5	45	90277	3	0.1	2	0	0	0	0	1	0
13	48	23	114	93106	2	3.8	3	0	0	1	0	0	0
14	59	32	40	94920	4	2.5	2	0	0	0	0	1	0
15	67	41	112	91741	1	2	1	0	0	1	0	0	0
.....													
4981	29	5	135	95762	3	5.3	1	0	1	0	1	1	1
4982	34	9	195	90266	2	3	1	122	0	0	0	1	0
4983	36	10	45	95126	4	0.2	1	0	0	0	0	0	1
4984	51	26	72	95370	1	2.9	1	0	0	0	0	0	0
4985	27	1	98	94043	4	2.3	3	0	0	0	0	0	1
4986	48	23	30	94720	3	1.7	2	162	0	0	0	1	0
4987	32	6	78	95825	1	2.9	3	0	0	0	0	0	0
4988	48	23	43	93943	3	1.7	2	159	0	0	0	1	0
4989	34	8	85	95134	1	2.5	1	136	0	0	0	0	1
4990	24	0	38	93555	1	1	3	0	0	0	0	1	0
4991	55	25	58	95023	4	2	3	219	0	0	0	0	1
4992	51	25	92	91330	1	1.9	2	100	0	0	0	0	1
4993	30	5	13	90037	4	0.5	3	0	0	0	0	0	0
4994	45	21	218	91801	2	6.67	1	0	0	0	0	1	0
4995	64	40	75	94588	3	2	3	0	0	0	0	1	0
4996	29	3	40	92697	1	1.9	3	0	0	0	0	1	0
4997	30	4	15	92037	4	0.4	1	85	0	0	0	1	0
4998	63	39	24	93023	2	0.3	3	0	0	0	0	0	0
4999	65	40	49	90034	3	0.5	2	0	0	0	0	1	0
5000	28	4	83	92612	3	0.8	1	0	0	0	0	1	1

Sumber: Bank Loan Approval - LR, DT, RF and AUC (kaggle.com)

Link dataset: https://bit.ly/datasetUASdatmin

Lampiran 2. Syntax Preprocessing

```
#PREPROCESSING DATA#

# Membaca data dan menghapus kolom ID
data_bankloan <- read.csv("C:/Lessons/SEM 6/bankloan.csv", sep = ',')
data_bankloan <- data_bankloan[ , !(names(data_bankloan) %in% c("ID"))]

#Melihat dimensi data_bankloan awal
dimensi_data_bankloan <- dim(data_bankloan)
dimensi_data_bankloan
# Melihat struktur data_bankloan
str(data_bankloan)

#Pindahkan kolom "Personal.Loan" ke kolom terakhir
data_bankloan <- data_bankloan[, c(setdiff(names(data_bankloan), "Personal.Loan"),
 "Personal.Loan")]

# Mengubah Personal.Loan menjadi faktor
data_bankloan$Personal.Loan <- as.factor(data_bankloan$Personal.Loan)

# 1. Mengecek  dam menghapus missing values
sum(is.na(data_bankloan))
# Menghapus baris dengan missing values
data_bankloan<- na.omit(data_bankloan)

# 2. Menghitung IQR dan menghilangkan outlier untuk beberapa kolom
remove_outliers <- function(data_bankloan, columns) {
  for (col in columns) {
    Q1 <- quantile(data_bankloan[[col]], 0.25, na.rm = TRUE)
    Q3 <- quantile(data_bankloan[[col]], 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1
    data_bankloan <- data_bankloan[!(data_bankloan[[col]] < (Q1 - 1.5 * IQR) |
 data_bankloan[[col]] > (Q3 + 1.5 * IQR)), ]
```

```
  }
  return(data_bankloan)
}
# Daftar kolom yang ingin dihilangkan outlier
columns_to_process <- c("Age","Experience", "Income", "ZIP.Code", "Family","CCAvg",
  "Education", "Mortgage","Securities.Account","CD.Account","Online", "CreditCard")
# Menghapus outlier dari kolom yang dipilih
data_bankloan <- remove_outliers(data_bankloan, columns_to_process)

dimensi_data_bankloan <- dim(data_bankloan)
dimensi_data_bankloan
data_bankloan

# 3. Menghitung standar deviasi untuk setiap kolom
sd_values <- apply(data_bankloan[, 1:12], 2, sd)
sd_values
# Identifikasi kolom yang memiliki standar deviasi nol
columns_to_remove_sd <- which(sd_values == 0)

# Periksa kolom yang akan dihapus
if(length(columns_to_remove_sd) > 0) {
  print("Kolom yang akan dihapus karena std = 0:")
  print(names(data_bankloan)[columns_to_remove_sd])
} else {
  print("Tidak ada kolom dengan standar deviasi nol.")
}

# Hapus kolom yang memiliki standar deviasi nol
if(length(columns_to_remove_sd) > 0) {
  data_bankloan_clean <- data_bankloan[, -columns_to_remove_sd]
} else {
  data_bankloan_clean <- data_bankloan
}
```

```
data_bankloan_clean
dim(data_bankloan_clean)

# 4. Menghapus Kolom Yang Sangat Berkorelasi
y=data_bankloan_clean[ , 11]
x=data_bankloan_clean[ , 1:10]
# Menghitung matriks korelasi
cor_matrix <- cor(x)
cor_matrix
# Identifikasi kolom yang sangat berkorelasi (korelasi > 0.99)
columns_to_remove_cor <- findCorrelation(cor_matrix, cutoff = 0.99)
# Periksa kolom yang akan dihapus
if(length(columns_to_remove_cor) > 0.99) {
  print("Kolom yang akan dihapus karena korelasi > 0.99:")
  print(names(data_bankloan)[columns_to_remove_cor])
} else {
  print("Tidak ada kolom yang sangat berkorelasi")
}

# Hapus kolom yang sangat berkorelasi
if(length(columns_to_remove_cor) > 0.99) {
  data_bankloan_clean <- data_bankloan_clean[, -columns_to_remove_cor]
} else {
  data_bankloan_clean <- data_bankloan_clean
}
data_bankloan_clean
dim(data_bankloan_clean)

# 5.Menghapus kolom yang memiliki IQR 0
# Menghitung nilai IQR untuk setiap kolom
iqr_values <- apply(data_bankloan_clean[, 1:9], 2, IQR)
iqr_values
# Identifikasi kolom yang memiliki IQR 0
```

```
columns_to_remove_iqr <- which(iqr_values == 0)
# Periksa kolom yang akan dihapus karena IQR = 0
if(length(columns_to_remove_iqr) > 0) {
  print("Kolom yang akan dihapus karena IQR = 0:")
  print(names(data_bankloan_clean)[columns_to_remove_iqr])
} else {
  print("Tidak ada kolom dengan IQR nol.")
}

# Hapus kolom yang memiliki IQR nol jika ada
if(length(columns_to_remove_iqr) > 0) {
  data_bankloan_clean <- data_bankloan_clean[, -columns_to_remove_iqr]
}
data_bankloan_clean
############################DEFINISI                                TRAINING
 TESTING##############################
#Mendefinisikan kembali data_bankloan dengan berisi variabel yang sudah di cleaning
data_bankloan<-data_bankloan_clean
set.seed(123)
#Membagi data menjadi training dan testing
indeks <- createDataPartition(data_bankloan$Personal.Loan, p = 0.7, list = FALSE)
data_training <- data_bankloan[indeks, ]
data_testing <- data_bankloan[-indeks, ]
dim(data_training)
dim(data_testing)
dim(data_bankloan)
```

Lampiran 3. Syntax Regresi logistik biner

```
##########################################UJI                                RASIO
LIKELIHOOD#######################################
# Bangun model lengkap
model_logistik_lengkap <- glm(Personal.Loan ~ ., data = data_training, family = binomial)
```

```
# Bangun model nol (sederhana)
model_logistik_nol <- glm(Personal.Loan ~ 1, data = data_training, family = binomial)

# Hitung nilai uji likelihood ratio
lrt_stat <- 2 * (logLik(model_logistik_lengkap) - logLik(model_logistik_nol))

# Hitung derajat kebebasan
df <- length(coef(model_logistik_lengkap)) - length(coef(model_logistik_nol))

# Hitung nilai p-value
p_value <- pchisq(lrt_stat, df, lower.tail = FALSE)

# Tampilkan hasil
cat("-------------------------Uji Rasio Likelihood----------------------------")
print(paste("Nilai uji likelihood ratio:", lrt_stat))
print(paste("Derajat kebebasan:", df))
print(paste("Nilai p-value:", p_value))

# Bandingkan dengan alpha (tingkat signifikansi)
alpha <- 0.05
if (p_value < alpha) {
  print("Tolak H0: Secara bersama-sama variabel bebas memengaruhi model")
} else {
  print("Terima H0: Secara bersama-sama variabel bebas tidak memengaruhi model")
}
cat("-----------------------------------------------------------------------")

##########################################UJI GOODNESS OF FIT#####################################
# Evaluasi Goodness of fit
predicted_values <- predict(model_logistik_lengkap, newdata = data_training, type = "response")
hoslem_test <- hoslem.test(data_training$Personal.Loan, predicted_values, g = 11)
```

```
# Tampilkan hasil
print("Uji Hosmer-Lemeshow Test:")
print(hoslem_test)

cat("-------------------------Uji Goodness of Fit----------------------------")
# Interpretasi hasil
cat("\nInterpretasi:\n")
if (hoslem_test$p.value < 0.05) {
  cat("Nilai p-value (", hoslem_test$p.value, ") < alpha (0.05).\n")
  cat("Tolak H0: Model tidak sesuai (observasi dan prediksi berbeda).\n")
} else {
  cat("Nilai p-value (", hoslem_test$p.value, ") >= alpha (0.05).\n")
  cat("Terima H0: Model sesuai (observasi dan prediksi tidak berbeda).\n")
}
cat("------------------------------------------------------------------------")

#########################################################UJI
WALD##############################################
# Mengambil estimasi koefisien
coef_est <- coef(model_logistik_lengkap)

# Mengambil kuadrat standar error
std_err <- summary(model_logistik_lengkap)$coefficients[, "Std. Error"]

# Menghitung statistik uji Wald
wald_stat <- (coef_est / std_err)^2

# Menghitung nilai p-value
p_value <- pchisq(wald_stat, df = 1, lower.tail = FALSE)  # df = 1 karena satu koefisien
yang diuji

cat("--------------------------------Uji Wald---------------------------------")
```

```
# Tampilkan hasil
print("Uji Wald Test:")
print("")

for (i in 1:length(coef_est)) {
  cat("Variabel:", names(coef_est)[i], "\n")
  cat("Estimasi Koefisien:", coef_est[i], "\n")
  cat("Kuadrat Standar Error:", std_err[i]^2, "\n")
  cat("Statistik uji Wald:", wald_stat[i], "\n")
  cat("Nilai p-value:", p_value[i], "\n")
  if (p_value[i] < 0.05) {
    cat("Tolak H0: Variabel bebas signifikan terhadap model\n\n")
  } else {
    cat("Terima H0: Variabel bebas tidak signifikan terhadap model\n\n")
  }
}

# Menyimpan nama variabel yang tidak signifikan
non_significant_vars <- names(coef_est)[p_value >= 0.05]

# Menampilkan variabel yang tidak signifikan
cat("Variabel yang tidak signifikan:\n")
print(non_significant_vars)

################################################## UJI RASIO LIKELIHOOD (PART 2) ##########################################
# Mengambil hanya variabel yang signifikan berdasarkan uji Wald
significant_vars <- names(coef_est)[p_value < 0.05]

# Exclude the Intercept term if it's present
significant_vars <- significant_vars[significant_vars != "(Intercept)"]

# Bangun model regresi logistik biner hanya dengan variabel yang signifikan
```

```
# Mengonversi significant_vars menjadi formula
formula_significant <- as.formula(paste("Personal.Loan ~", paste(significant_vars, collapse = "+")))
model_logistik_significant <- glm(formula_significant, data = data_training, family = binomial)

# Bangun model nol (sederhana)
model_logistik_nol_significant <- glm(Personal.Loan ~ 1, data = data_training, family = binomial)

# Hitung nilai uji likelihood ratio
lrt_stat_significant <- 2 * (logLik(model_logistik_significant) - logLik(model_logistik_nol_significant))

# Hitung derajat kebebasan
df_significant <- length(coef(model_logistik_significant)) - length(coef(model_logistik_nol_significant))

# Hitung nilai p-value
p_value_significant <- pchisq(lrt_stat_significant, df_significant, lower.tail = FALSE)

# Tampilkan hasil uji rasio likelihood
cat("-------------------------Uji Rasio Likelihood (Variabel Signifikan)----------------------------")
print(paste("Nilai uji likelihood ratio:", lrt_stat_significant))
print(paste("Derajat kebebasan:", df_significant))
print(paste("Nilai p-value:", p_value_significant))

# Bandingkan dengan alpha (tingkat signifikansi)
alpha <- 0.05
if (p_value_significant < alpha) {
  print("Tolak H0: Secara bersama-sama variabel bebas yang signifikan memengaruhi model")
```

```
} else {
  print("Terima H0: Secara bersama-sama variabel bebas yang signifikan tidak memengaruhi model")
}
cat("-----------------------------------------------------------------------------------------")

##################################################UJI GOODNESS OF FIT (PART 2)##########################################
# Evaluasi Goodness of fit hanya dengan variabel yang signifikan
hoslem_test_significant <- hoslem.test(data_training$Personal.Loan, fitted(model_logistik_significant), g = 7)

# Tampilkan hasil
cat("-------------------------Uji Goodness of Fit (Variabel Signifikan)-------------------------")
print("Uji Hosmer-Lemeshow Test:")
print(hoslem_test_significant)

# Interpretasi hasil
cat("\nInterpretasi:\n")
if (hoslem_test_significant$p.value < 0.05) {
  cat("Nilai p-value (", hoslem_test_significant$p.value, ") < alpha (0.05).\n")
  cat("Tolak H0: Model tidak sesuai (observasi dan prediksi berbeda).\n")
} else {
  cat("Nilai p-value (", hoslem_test_significant$p.value, ") >= alpha (0.05).\n")
  cat("Terima H0: Model sesuai (observasi dan prediksi tidak berbeda).\n")
}
cat("-----------------------------------------------------------------------------------------")

##################################################UJI WALD (PART 2)##########################################
# Mengambil estimasi koefisien hanya untuk variabel yang signifikan
coef_est_significant <- coef(model_logistik_significant)
```

```
# Mengambil kuadrat standar error hanya untuk variabel yang signifikan
std_err_significant <- summary(model_logistik_significant)$coefficients[, "Std. Error"]

# Menghitung statistik uji Wald
wald_stat_significant <- (coef_est_significant / std_err_significant)^2

# Menampilkan panjang vektor
#cat("Panjang vektor coef_est_significant:", length(coef_est_significant), "\n")
#cat("Panjang vektor std_err_significant:", length(std_err_significant), "\n")

# Memeriksa variabel yang mungkin menyebabkan perbedaan panjang
#cat("Variabel yang mungkin menyebabkan perbedaan panjang:\n")
#print(setdiff(names(model_logistik_significant$coefficients),
names(coef_est_significant)))
#print(setdiff(names(model_logistik_significant$coefficients),
names(std_err_significant)))



# Menghitung nilai p-value
p_value_significant <- pchisq(wald_stat_significant, df = 1, lower.tail = FALSE)  # df = 1
karena satu koefisien yang diuji

cat("--------------------------------Uji Wald (Variabel Signifikan)----------------------------------
")
# Tampilkan hasil
print("Uji Wald Test:")
print("")

for (i in 1:length(coef_est_significant)) {
  cat("Variabel:", names(coef_est_significant)[i], "\n")
  cat("Estimasi Koefisien:", coef_est_significant[i], "\n")
  cat("Kuadrat Standar Error:", std_err_significant[i]^2, "\n")
```

```
  cat("Statistik uji Wald:", wald_stat_significant[i], "\n")
  cat("Nilai p-value:", p_value_significant[i], "\n")
  if (p_value_significant[i] < 0.05) {
    cat("Tolak H0: Variabel bebas signifikan terhadap model\n\n")
  } else {
    cat("Terima H0: Variabel bebas tidak signifikan terhadap model\n\n")
  }
}

##################################################MODEL
AKHIRRR##########################################
# Mengonversi significant_vars menjadi formula
formula_significant <- as.formula(paste("Personal.Loan ~", paste(significant_vars,
collapse = "+")))
model_logistik_significant <- glm(formula_significant, data = data_training, family =
binomial)

# Evaluasi Multikolineritas
vif_values <- vif(model_logistik_significant)
print(vif_values)

# Prediksi pada data testing
predictions <- predict(model_logistik_significant, newdata = data_testing, type =
"response")

# Ubah prediksi menjadi kelas biner menggunakan threshold 0.5
predicted_classes <- ifelse(predictions >= 0.5, 1, 0)

# Hitung matriks kebingungan
confusion_matrix <- table(data_testing$Personal.Loan, predicted_classes)

# Tampilkan Confusion Matrix
print("Confusion Matrix:")
```

```
print(confusion_matrix)

# Hitung nilai precision, recall, dan F1 score
TP <- confusion_matrix[2, 2]
FP <- confusion_matrix[1, 2]
FN <- confusion_matrix[2, 1]
TN <- confusion_matrix[1, 1]

precision <- TP / (TP + FP)
recall <- TP / (TP + FN)
f1_score <- 2 * (precision * recall) / (precision + recall)

# Hitung akurasi dalam persentase
accuracy <- (TP + TN) / sum(confusion_matrix) * 100

# Tampilkan nilai precision, recall, F1 score, dan akurasi dalam persentase
print(paste("Precision:", precision))
print(paste("Recall:", recall))
print(paste("F1 Score:", f1_score))
print(paste("Accuracy:", accuracy, "%"))
```

Lampiran 4. Syntax Analisis Diskriminan

```
########ANALISIS DISKRIMINAN PADA DATA BANKLOAN########

#Spesifikasi Variabel
y=data_training[ , 10]
x=data_training[ , 1:9]

###UJI ASUMSI###
#UJI NORMALITAS MULTIVARIAT
# Lakukan uji normalitas multivariat secara formal
normality_test <- mvn(x, mvnTest = "mardia")
print(normality_test)
```

```
# Menampilkan Q-Q plot multivariat untuk memeriksa normalitas multivariat
hasildata <- mvn(data = x, multivariatePlot = 'qq')

# UJI HOMOGENITAS KOVARIAN
box_m_test <- boxM(data=x, group=y)
print(box_m_test)

#UJI WILK LAMBDA
m <- manova(formula = cbind(data_training$Experience, data_training$Income, data_training$ZIP.Code,
                    data_training$Family, data_training$CCAvg, data_training$Education, data_training$Mortgage, data_training$Online,
                    data_training$CreditCard) ~ data_training$Personal.Loan)
summary(object = m, test = 'Wilks')

#UJI NON MULTIKOLINIERITAS
VIF=function(x){
  VIF=diag(solve(cor(x)))
  result=ifelse(VIF>10,"mulicolinearity", "non multicolinearity")
  data1=data.frame(VIF,result)
  return(data1)
}
VIF(x)

#Analisis Diskriminan
linearDA <- lda(formula = Personal.Loan ~., data = data_training)
linearDA

plot(linearDA, col = as.integer(data_training$Personal.Loan))

# Melakukan prediksi
predicted <- predict(object = linearDA, newdata = data_testing)
# Pastikan prediksi dilakukan pada data testing yang benar
```

```
predicted <- predict(linearDA, data_testing)

# Memastikan panjang data pengujian dan prediksi sama
print(paste("Panjang data pengujian:", nrow(data_testing)))
print(paste("Panjang prediksi:", length(predicted$class)))

# Menghitung confusion matrix jika panjang vektor sama
if (nrow(data_testing) == length(predicted$class)) {
  conf_matrix <- table(actual = data_testing$Personal.Loan, predicted = predicted$class)
  print(conf_matrix)
} else {
  stop("Panjang data pengujian dan prediksi tidak sama.")
}

# Menghitung akurasi model
accuracy <- sum(predicted$class == data_testing$Personal.Loan) / nrow(data_testing)

# Menghitung precision, recall, dan f1-score
precision <- posPredValue(conf_matrix, positive = "1")
recall <- sensitivity(conf_matrix, positive = "1")
f1 <- 2 * (precision * recall) / (precision + recall)

# Menampilkan hasil accuracy, precision, recall, dan f1-score
print(paste("Akurasi:", accuracy))
print(paste("Precision:", precision))
print(paste("Recall:", recall))
print(paste("F1-Score:", f1))
```

Lampiran 5. *Output* Regresi logistik biner

```
> cat("-------------------------Uji Rasio Likelihood-------------------
---------")
-------------------------Uji Rasio Likelihood-------------------------
> print(paste("Nilai uji likelihood ratio:", lrt_stat))
[1] "Nilai uji likelihood ratio: 615.514996961772"
> print(paste("Derajat kebebasan:", df))
[1] "Derajat kebebasan: 9"
> print(paste("Nilai p-value:", p_value))
```

```
[1] "Nilai p-value: 9.78709233437283e-127"
> # Bandingkan dengan alpha (tingkat signifikansi)
> alpha <- 0.05
> if (p_value < alpha) {
+   print("Tolak H0: Secara bersama-sama variabel bebas memengaruhi mode
l")
+ } else {
+   print("Terima H0: Secara bersama-sama variabel bebas tidak memengaru
hi model")
+ }
[1] "Tolak H0: Secara bersama-sama variabel bebas memengaruhi model"
> cat("-------------------------------------------------------------------
---------")
--------------------------------------------------------------------------
-->
> ##########################################UJI GOODNESS OF FIT##########
############################
> # Evaluasi Goodness of fit
> predicted_values <- predict(model_logistik_lengkap, newdata = data_tra
ining, type = "response")
> hoslem_test <- hoslem.test(data_training$Personal.Loan, predicted_valu
es, g = 11)
>
> # Tampilkan hasil
> print("Uji Hosmer-Lemeshow Test:")
[1] "Uji Hosmer-Lemeshow Test:"
> print(hoslem_test)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  data_training$Personal.Loan, predicted_values
X-squared = 9.3524, df = 9, p-value = 0.4054

>
> cat("-------------------------Uji Goodness of Fit--------------------
--------")
-------------------------Uji Goodness of Fit--------------------------
-> # Interpretasi hasil
> cat("\nInterpretasi:\n")

Interpretasi:
> if (hoslem_test$p.value < 0.05) {
+   cat("Nilai p-value (", hoslem_test$p.value, ") < alpha (0.05).\n")
+   cat("Tolak H0: Model tidak sesuai (observasi dan prediksi berbeda).\
n")
+ } else {
+   cat("Nilai p-value (", hoslem_test$p.value, ") >= alpha (0.05).\n")
+   cat("Terima H0: Model sesuai (observasi dan prediksi tidak berbeda).
\n")
+ }
Nilai p-value ( 0.4053978 ) >= alpha (0.05).
Terima H0: Model sesuai (observasi dan prediksi tidak berbeda).
> cat("-------------------------------------------------------------------
---------")
--------------------------------------------------------------------------
-->
> ##################################################UJI WALD#############
############################
> # Mengambil estimasi koefisien
> coef_est <- coef(model_logistik_lengkap)
> # Mengambil kuadrat standar error
> std_err <- summary(model_logistik_lengkap)$coefficients[, "Std. Error"
]
>
> # Menghitung statistik uji Wald
> wald_stat <- (coef_est / std_err)^2
> # Menghitung nilai p-value
> p_value <- pchisq(wald_stat, df = 1, lower.tail = FALSE)  # df = 1 kar
ena satu koefisien yang diuji
```

```
> cat("--------------------------------Uji Wald-------------------------
---------")
--------------------------------Uji Wald--------------------------------
> # Tampilkan hasil
> print("Uji Wald Test:")
[1] "Uji Wald Test:"
> print("")
[1] ""
>
> for (i in 1:length(coef_est)) {
+   cat("Variabel:", names(coef_est)[i], "\n")
+   cat("Estimasi Koefisien:", coef_est[i], "\n")
+   cat("Kuadrat Standar Error:", std_err[i]^2, "\n")
+   cat("Statistik uji Wald:", wald_stat[i], "\n")
+   cat("Nilai p-value:", p_value[i], "\n")
+   if (p_value[i] < 0.05) {
+     cat("Tolak H0: Variabel bebas signifikan terhadap model\n\n")
+   } else {
+     cat("Terima H0: Variabel bebas tidak signifikan terhadap model\n\n
")
+   }
+ }
Variabel: (Intercept)
Estimasi Koefisien: -13.21305
Kuadrat Standar Error: 40.80414
Statistik uji Wald: 4.278602
Nilai p-value: 0.03859509
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: Experience
Estimasi Koefisien: -9.904785e-05
Kuadrat Standar Error: 0.0001116868
Statistik uji Wald: 8.783916e-05
Nilai p-value: 0.9925221
Terima H0: Variabel bebas tidak signifikan terhadap model

Variabel: Income
Estimasi Koefisien: 0.06411192
Kuadrat Standar Error: 2.351214e-05
Statistik uji Wald: 174.8177
Nilai p-value: 6.561574e-40
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: ZIP.Code
Estimasi Koefisien: -1.822512e-05
Kuadrat Standar Error: 4.596299e-09
Statistik uji Wald: 0.07226572
Nilai p-value: 0.788066
Terima H0: Variabel bebas tidak signifikan terhadap model

Variabel: Family
Estimasi Koefisien: 0.5437861
Kuadrat Standar Error: 0.01462142
Statistik uji Wald: 20.22398
Nilai p-value: 6.888386e-06
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: CCAvg
Estimasi Koefisien: 0.6212445
Kuadrat Standar Error: 0.009210496
Statistik uji Wald: 41.90271
Nilai p-value: 9.592944e-11
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: Education
Estimasi Koefisien: 1.622713
Kuadrat Standar Error: 0.03288499
Statistik uji Wald: 80.07298
Nilai p-value: 3.608328e-19
```

```
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: Mortgage
Estimasi Koefisien: -0.0009187861
Kuadrat Standar Error: 3.763947e-06
Statistik uji Wald: 0.2242773
Nilai p-value: 0.6358
Terima H0: Variabel bebas tidak signifikan terhadap model

Variabel: Online
Estimasi Koefisien: -0.33622
Kuadrat Standar Error: 0.06180417
Statistik uji Wald: 1.829066
Nilai p-value: 0.176238
Terima H0: Variabel bebas tidak signifikan terhadap model

Variabel: CreditCard
Estimasi Koefisien: -0.6791333
Kuadrat Standar Error: 0.08966303
Statistik uji Wald: 5.143948
Nilai p-value: 0.02332753
Tolak H0: Variabel bebas signifikan terhadap model

> # Menyimpan nama variabel yang tidak signifikan
> non_significant_vars <- names(coef_est)[p_value >= 0.05]
> # Menampilkan variabel yang tidak signifikan
> cat("Variabel yang tidak signifikan:\n")
Variabel yang tidak signifikan:
> print(non_significant_vars)
[1] "Experience" "ZIP.Code"   "Mortgage"   "Online"
> ################################################ UJI RASIO LIKELIHOOD
(PART 2) ##########################################
> # Mengambil hanya variabel yang signifikan berdasarkan uji Wald
> significant_vars <- names(coef_est)[p_value < 0.05]
> # Exclude the Intercept term if it's present
> significant_vars <- significant_vars[significant_vars != "(Intercept)"
]
> # Bangun model regresi logistik biner hanya dengan variabel yang signi
fikan
> # Mengonversi significant_vars menjadi formula
> formula_significant <- as.formula(paste("Personal.Loan ~", paste(signi
ficant_vars, collapse = "+")))
> model_logistik_significant <- glm(formula_significant, data = data_tra
ining, family = binomial)
> # Bangun model nol (sederhana)
> model_logistik_nol_significant <- glm(Personal.Loan ~ 1, data = data_t
raining, family = binomial)
> # Hitung nilai uji likelihood ratio
> lrt_stat_significant <- 2 * (logLik(model_logistik_significant) - logL
ik(model_logistik_nol_significant))
> # Hitung derajat kebebasan
> df_significant <- length(coef(model_logistik_significant)) - length(co
ef(model_logistik_nol_significant))
> # Hitung nilai p-value
> p_value_significant <- pchisq(lrt_stat_significant, df_significant, lo
wer.tail = FALSE)
> # Tampilkan hasil uji rasio likelihood
> cat("--------------------------Uji Rasio Likelihood (Variabel Signifik
an)----------------------------")
--------------------------Uji Rasio Likelihood (Variabel Signifikan)----
------------------------
> print(paste("Nilai uji likelihood ratio:", lrt_stat_significant))
[1] "Nilai uji likelihood ratio: 613.344689820352"
> print(paste("Derajat kebebasan:", df_significant))
[1] "Derajat kebebasan: 5"
> print(paste("Nilai p-value:", p_value_significant))
[1] "Nilai p-value: 2.64478920899501e-130"
> # Bandingkan dengan alpha (tingkat signifikansi)
> alpha <- 0.05
```

```
> if (p_value_significant < alpha) {
+   print("Tolak H0: Secara bersama-sama variabel bebas yang signifikan
memengaruhi model")
+ } else {
+   print("Terima H0: Secara bersama-sama variabel bebas yang signifikan
tidak memengaruhi model")
+ }
[1] "Tolak H0: Secara bersama-sama variabel bebas yang signifikan memeng
aruhi model"
> cat("------------------------------------------------------------------
--------------------------------")
------------------------------------------------------------------------
------------------------
> ################################################UJI GOODNESS OF FIT (
PART 2)#########################################
> # Evaluasi Goodness of fit hanya dengan variabel yang signifikan
> hoslem_test_significant <- hoslem.test(data_training$Personal.Loan, fi
tted(model_logistik_significant), g = 7)
> # Tampilkan hasil
> cat("-------------------------Uji Goodness of Fit (Variabel Signifika
n)----------------------------")
-------------------------Uji Goodness of Fit (Variabel Signifikan)-----
----------------------> print("Uji Hosmer-Lemeshow Test:")
[1] "Uji Hosmer-Lemeshow Test:"
> print(hoslem_test_significant)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  data_training$Personal.Loan, fitted(model_logistik_significant)
X-squared = 10.288, df = 5, p-value = 0.06746

> # Interpretasi hasil
> cat("\nInterpretasi:\n")

Interpretasi:
> if (hoslem_test_significant$p.value < 0.05) {
+   cat("Nilai p-value (", hoslem_test_significant$p.value, ") < alpha (
0.05).\n")
+   cat("Tolak H0: Model tidak sesuai (observasi dan prediksi berbeda).\
n")
+ } else {
+   cat("Nilai p-value (", hoslem_test_significant$p.value, ") >= alpha
(0.05).\n")
+   cat("Terima H0: Model sesuai (observasi dan prediksi tidak berbeda).
\n")
+ }
Nilai p-value ( 0.06746457 ) >= alpha (0.05).
Terima H0: Model sesuai (observasi dan prediksi tidak berbeda).
> cat("------------------------------------------------------------------
--------------------------------")
------------------------------------------------------------------------
------------------------
> ################################################UJI WALD (PART 2)####
#####################################
> # Mengambil estimasi koefisien hanya untuk variabel yang signifikan
> coef_est_significant <- coef(model_logistik_significant)
> # Mengambil kuadrat standar error hanya untuk variabel yang signifikan
> std_err_significant <- summary(model_logistik_significant)$coefficient
s[, "Std. Error"]
> # Menghitung statistik uji Wald
> wald_stat_significant <- (coef_est_significant / std_err_significant)^
2
> # Menampilkan panjang vektor
> #cat("Panjang vektor coef_est_significant:", length(coef_est_significa
nt), "\n")
> #cat("Panjang vektor std_err_significant:", length(std_err_significant
), "\n")
> # Memeriksa variabel yang mungkin menyebabkan perbedaan panjang
> #cat("Variabel yang mungkin menyebabkan perbedaan panjang:\n")
```

```
> #print(setdiff(names(model_logistik_significant$coefficients), names(c
oef_est_significant)))
> #print(setdiff(names(model_logistik_significant$coefficients), names(s
td_err_significant)))
> # Menghitung nilai p-value
> p_value_significant <- pchisq(wald_stat_significant, df = 1, lower.tai
l = FALSE)  # df = 1 karena satu koefisien yang diuji
> cat("--------------------------------Uji Wald (Variabel Signifikan)---
-------------------------------")
--------------------------------Uji Wald (Variabel Signifikan)----------
------------------------> # Tampilkan hasil
> print("Uji Wald Test:")
[1] "Uji Wald Test:"
> print("")
[1] ""
> for (i in 1:length(coef_est_significant)) {
+   cat("Variabel:", names(coef_est_significant)[i], "\n")
+   cat("Estimasi Koefisien:", coef_est_significant[i], "\n")
+   cat("Kuadrat Standar Error:", std_err_significant[i]^2, "\n")
+   cat("Statistik uji Wald:", wald_stat_significant[i], "\n")
+   cat("Nilai p-value:", p_value_significant[i], "\n")
+   if (p_value_significant[i] < 0.05) {
+     cat("Tolak H0: Variabel bebas signifikan terhadap model\n\n")
+   } else {
+     cat("Terima H0: Variabel bebas tidak signifikan terhadap model\n\n
")
+   }
+ }
Variabel: (Intercept)
Estimasi Koefisien: -15.1792
Kuadrat Standar Error: 1.005846
Statistik uji Wald: 229.0689
Nilai p-value: 9.514793e-52
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: Income
Estimasi Koefisien: 0.064439
Kuadrat Standar Error: 2.338253e-05
Statistik uji Wald: 177.5849
Nilai p-value: 1.632125e-40
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: Family
Estimasi Koefisien: 0.5406403
Kuadrat Standar Error: 0.0142843
Statistik uji Wald: 20.46246
Nilai p-value: 6.081242e-06
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: CCAvg
Estimasi Koefisien: 0.61929
Kuadrat Standar Error: 0.009104597
Statistik uji Wald: 42.12378
Nilai p-value: 8.567483e-11
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: Education
Estimasi Koefisien: 1.63328
Kuadrat Standar Error: 0.03294774
Statistik uji Wald: 80.96471
Nilai p-value: 2.297849e-19
Tolak H0: Variabel bebas signifikan terhadap model

Variabel: CreditCard
Estimasi Koefisien: -0.5936488
Kuadrat Standar Error: 0.08397471
Statistik uji Wald: 4.196726
Nilai p-value: 0.0405021
Tolak H0: Variabel bebas signifikan terhadap model
```

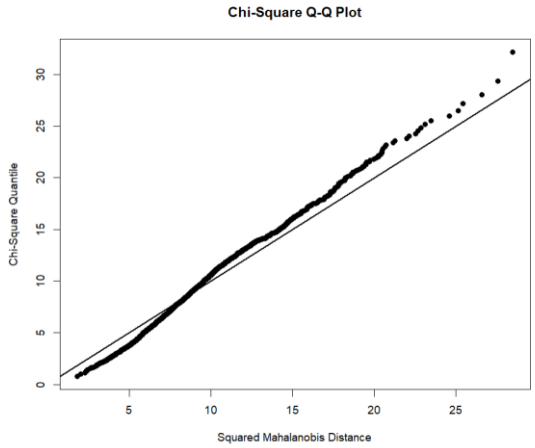
```
> ##############################################MODEL AKHIRRR########
##################################
> # Mengonversi significant_vars menjadi formula
> formula_significant <- as.formula(paste("Personal.Loan ~", paste(signi
ficant_vars, collapse = "+")))
> model_logistik_significant <- glm(formula_significant, data = data_tra
ining, family = binomial)
> # Evaluasi Multikolineritas
> vif_values <- vif(model_logistik_significant)
> print(vif_values)
    Income     Family      CCAvg  Education CreditCard
  1.941874   1.387365   1.053960   1.715213   1.006583
> # Prediksi pada data testing
> predictions <- predict(model_logistik_significant, newdata = data_test
ing, type = "response")
> # Ubah prediksi menjadi kelas biner menggunakan threshold 0.5
> predicted_classes <- ifelse(predictions >= 0.5, 1, 0)
> # Hitung matriks kebingungan
> confusion_matrix <- table(data_testing$Personal.Loan, predicted_classe
s)
> # Tampilkan Confusion Matrix
> print("Confusion Matrix:")
[1] "Confusion Matrix:"
> print(confusion_matrix)
   predicted_classes
       0    1
  0 1080   10
  1   25   31
>
> # Hitung nilai precision, recall, dan F1 score
> TP <- confusion_matrix[2, 2]
> FP <- confusion_matrix[1, 2]
> FN <- confusion_matrix[2, 1]
> TN <- confusion_matrix[1, 1]
> precision <- TP / (TP + FP)
> recall <- TP / (TP + FN)
> f1_score <- 2 * (precision * recall) / (precision + recall)
> # Hitung akurasi dalam persentase
> accuracy <- (TP + TN) / sum(confusion_matrix) * 100
> # Tampilkan nilai precision, recall, F1 score, dan akurasi dalam perse
ntase
> print(paste("Precision:", precision))
[1] "Precision: 0.75609756097561"
> print(paste("Recall:", recall))
[1] "Recall: 0.553571428571429"
> print(paste("F1 Score:", f1_score))
[1] "F1 Score: 0.639175257731959"
> print(paste("Accuracy:", accuracy, "%"))
[1] "Accuracy: 96.9458987783595 %"
```

Lampiran 6. *Output* Analisis Diskriminan

```
> #UJI NORMALITAS MULTIVARIAT
> # Lakukan uji normalitas multivariat secara formal
> normality_test <- mvn(x, mvnTest = "mardia")
> print(normality_test)
$multivariateNormality
            Test         Statistic p value Result
1 Mardia Skewness  2935.08539741604       0     NO
2 Mardia Kurtosis -10.9231661986839       0     NO
3             MVN              <NA>    <NA>     NO

$univariateNormality
              Test   Variable Statistic   p value Normality
1 Anderson-Darling Experience   24.9096  <0.001      NO
2 Anderson-Darling   Income     45.1623  <0.001      NO
3 Anderson-Darling  ZIP.Code    50.9259  <0.001      NO
```

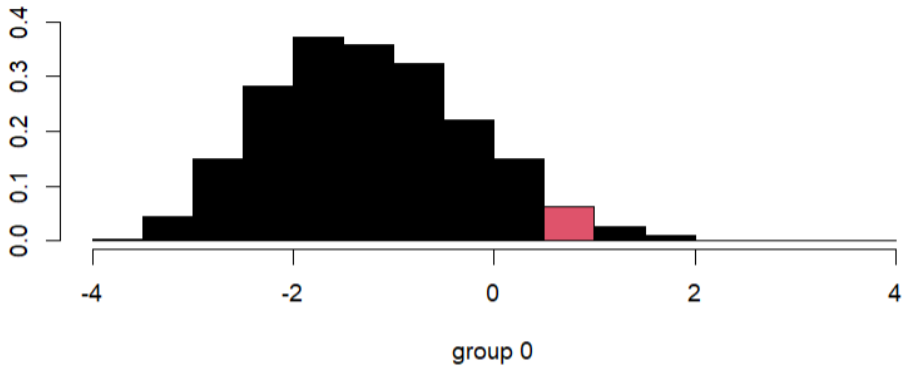
```
4 Anderson-Darling    Family    151.3774  <0.001       NO
5 Anderson-Darling    CCAvg      41.3559  <0.001       NO
6 Anderson-Darling Education    239.6579  <0.001       NO
7 Anderson-Darling  Mortgage    493.9665  <0.001       NO
8 Anderson-Darling    Online    497.9487  <0.001       NO
9 Anderson-Darling CreditCard   604.9057  <0.001       NO

$Descriptives
                n          Mean       Std.Dev  Median    Min      Max     25th
75th         Skew    Kurtosis
Experience 2676 2.035164e+01   11.4562818    21.0     -3    43.0     10.0
30.0 -0.06843224 -1.1188066
Income     2676 6.312108e+01   37.3050610    55.0      8   185.0     34.0
83.0  0.92737416  0.5085942
ZIP.Code   2676 9.314780e+04 1785.8110204 93407.0 90005 96651.0 91768.0
94609.0 -0.26816514 -1.1373523
Family     2676 2.420777e+00    1.1626918     2.0      1     4.0      1.0
4.0  0.11074081 -1.4482725
CCAvg      2676 1.538864e+00    1.1381331     1.4      0     5.2      0.6
2.2  0.87763492  0.4097559
Education  2676 1.936472e+00    0.8448194     2.0      1     3.0      1.0
3.0  0.12057219 -1.5891708
Mortgage   2676 3.782885e+01   66.5716833     0.0      0   245.0      0.0
83.0  1.48878290  0.8565085
Online     2676 5.822123e-01    0.4932870     1.0      0     1.0      0.0
1.0 -0.33319967 -1.8896837
CreditCard 2676 2.806428e-01    0.4493972     0.0      0     1.0      0.0
1.0  0.97586396 -1.0480808
```

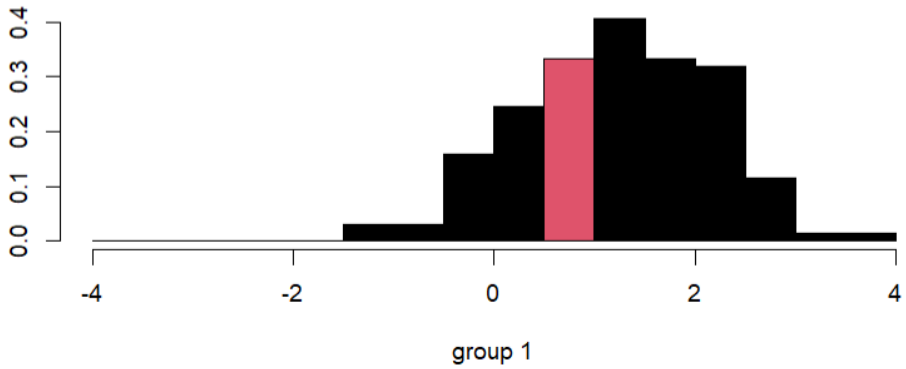
```
> # Menampilkan Q-Q plot multivariat untuk memeriksa normalitas multivar
iat
> hasildata <- mvn(data = x, multivariatePlot = 'qq')
```




```
> # UJI HOMOGENITAS KOVARIAN
> box_m_test <- boxM(data=x, group=y)
> print(box_m_test)

        Box's M-test for Homogeneity of Covariance Matrices

data:  x
Chi-Sq (approx.) = 171.63, df = 45, p-value < 2.2e-16

> #UJI WILK LAMBDA
> m <- manova(formula = cbind(data_training$Experience, data_training$In
come, data_training$ZIP.Code,
+                             data_training$Family, data_training$CCAvg,
data_training$Education, data_training$Mortgage, data_training$Online,
+                             data_training$CreditCard) ~ data_training$
Personal.Loan)
> summary(object = m, test = 'Wilks')
                             Df   Wilks approx F num Df den Df    Pr(>F
)
```

```
data_training$Personal.Loan    1 0.77139   87.789      9   2666 < 2.2e-1
6 ***
Residuals                   2674
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #UJI NON MULTIKOLINIERITAS
> VIF=function(x){
+   VIF=diag(solve(cor(x)))
+   result=ifelse(VIF>10,"mulicolinearity", "non multicolinearity")
+   data1=data.frame(VIF,result)
+   return(data1)
+ }
> VIF(x)
                VIF                result
Experience 1.006973 non multicolinearity
Income     1.437899 non multicolinearity
ZIP.Code   1.005585 non multicolinearity
Family     1.032681 non multicolinearity
CCAvg      1.360088 non multicolinearity
Education  1.037378 non multicolinearity
Mortgage   1.009770 non multicolinearity
Online     1.005001 non multicolinearity
CreditCard 1.004353 non multicolinearity
> #Analisis Diskriminan
> linearDA <- lda(formula = Personal.Loan ~., data = data_training)
> linearDA
Call:
lda(Personal.Loan ~ ., data = data_training)

Prior probabilities of groups:
         0          1
0.94843049 0.05156951

Group means:
  Experience    Income ZIP.Code   Family    CCAvg Education Mortgage
Online CreditCard
0   20.35461  59.48897 93142.88 2.417652 1.456198  1.916470 38.35894 0.5
878645   0.286446
1   20.29710 129.92029 93238.24 2.478261 3.059203  2.304348 28.07971 0.4
782609   0.173913

Coefficients of linear discriminants:
                     LD1
Experience  3.531490e-03
Income      2.541651e-02
ZIP.Code    5.042881e-05
Family      1.526028e-01
CCAvg       2.602003e-01
Education   5.009759e-01
Mortgage    7.440088e-05
Online     -1.819479e-01
CreditCard -2.317133e-01
> plot(linearDA, col = as.integer(data_training$Personal.Loan))
```






```
> # Menghitung confusion matrix jika panjang vektor sama
> if (nrow(data_testing) == length(predicted$class)) {
+   conf_matrix <- table(actual = data_testing$Personal.Loan, predicted
= predicted$class)
+   print(conf_matrix)
+ } else {
+   stop("Panjang data pengujian dan prediksi tidak sama.")
+ }
      predicted
actual    0    1
     0 1074   13
     1   27   31
> # Menghitung akurasi model
> accuracy <- sum(predicted$class == data_testing$Personal.Loan) / nrow(
data_testing)
> # Menghitung precision, recall, dan f1-score
> precision <- posPredValue(conf_matrix, positive = "1")
> recall <- sensitivity(conf_matrix, positive = "1")
> f1 <- 2 * (precision * recall) / (precision + recall)
> # Menampilkan hasil accuracy, precision, recall, dan f1-score
> print(paste("Akurasi:", accuracy))
[1] "Akurasi: 0.965065502183406"
> print(paste("Precision:", precision))
[1] "Precision: 0.534482758620689"
> print(paste("Recall:", recall))
[1] "Recall: 0.704545454545455"
> print(paste("F1-Score:", f1))
[1] "F1-Score: 0.607843137254902"
```